A Review on New Technologies in 3GPP Standards for 5G Access and Beyond

Nhu-Ngoc Dao[a], Ngo Hoang Tu[b,c], Trong-Dai Hoang[d], Tri-Hai Nguyen[e], Luong Vuong Nguyen[f], Kyungchun Lee[g], Laihyuk Park[e], Woongsoo Na[h], Sungrae Cho[i]

[a]*Department of Computer Science and Engineering, Sejong University, Seoul 05006, South Korea*
[b]*Department of Electrical and Information Engineering, Seoul National University of Science and Technology, Seoul 01811, South Korea*
[c]*Department of Computer Engineering, Ho Chi Minh City University of Transport, Ho Chi Minh City 710372, Vietnam*
[d]*School of Electrical and Data Engineering, University of Technology Sydney, Ultimo, NSW 2007, Australia*
[e]*Department of Computer Science and Engineering, Seoul National University of Science and Technology, Seoul 01811, South Korea*
[f]*Department of Artificial Intelligence, FPT University, Danang 550000, Vietnam*
[g]*Department of Electrical and Information Engineering and the Research Center for Electrical and Information Technology, Seoul National University of Science and Technology, Seoul 01811, South Korea*
[h]*Department of Computer Science and Engineering, Kongju National University, Cheonan 31080, South Korea*
[i]*School of Computer Science and Engineering, Chung-Ang University, Seoul 06974, South Korea*

Abstract

The world is witnessing the rapid development of commercial fifth-generation (5G) networks that enable diverse applications and services with enhanced mobile broadband, ultra-reliability, low latency, and massive connectivity performance. Technical specifications and requirements for 5G networks have been officially standardized by the Third Generation Partnership Project (3GPP) organization since Release (Rel) 15. Subsequently, several advanced technologies have been involved in the standards through recent updates. This paper investigated the evolution of 5G access networks derived from stable 3GPP standards to obtain a systematic view of the technology. In particular, Rel 15 initiated the first complete 5G network along with the standalone New Radio (NR) architecture. While Rel 16 focused on improving system performance, expanding user cases, and exploiting new spectrum resources, Rel 17 targeted the cellular Internet of Things, multimedia services, and intelligent computing capabilities. In addition, expected features in Rel 18 and Rel 19 are studied and considered for future research for 5G and beyond. This paper aims to provide interested readers with a reference framework on the standard 5G access technology development and trends.

Keywords: B5G, communication standard, mobile network, wireless communication, access network, 3GPP standard

1. Introduction

Fifth-generation (5G) mobile technology has successfully proved its valuable contribution to societal development. This technology expands the capacities of existing services with extremely high performance while enabling new classes of applications based on three major use cases: enhanced mobile broadband (eMBB), ultra-reliable and low-latency communications (URLLC), and massive machine-type communications (mMTC). Indeed, the regular mobility report [1] published by Ericsson in November 2023 revealed that more than 280 service providers have commercially launched 5G services globally. As a result, global 5G mobile subscriptions are projected to reach 1.6 billion by the end of 2023. That number is forecasted to surpass 5.3 billion in 2029. Myriad 5G applications, such as mobile Internet broadband, real-time virtual reality, high-precision smart factories, heterogeneous smart cities and transportation, and aerial services, are offered.

The International Telecommunication Union Radiocommunication (ITU-R) sector has developed and managed key performance indicators (KPIs) and technical specifications for 5G recognition globally to achieve such rapid development. Detailed requirements for key metrics, such as the peak data rate, latency, mobility, connectivity, and spectral efficiency, are described in the latest standard recommendation ITU-R M.2150-1 [2]. Accordingly, the Third Generation Partnership Project (3GPP) organization has integrated appropriate technologies to enable 5G networks, devices, and services to satisfy the ITU-R requirements in their release standards, starting from Release (Rel) 15 [3].

In late 2018, 3GPP Rel 15 was completed, defining the first stage of 5G (i.e., 5G basic). In addition, Rel 15 completed 4G with the most advanced technology, Long-Term Evolution-Advanced (LTE-A) Pro, to ensure a smooth transition between two contiguous generations. The standard started 5G by intro-

Email addresses: nndao@sejong.ac.kr (Nhu-Ngoc Dao), tu.ngo@ut.edu.vn (Ngo Hoang Tu), dai.t.hoang@student.uts.edu.au (Trong-Dai Hoang), haint93@seoultech.ac.kr (Tri-Hai Nguyen), vuongnl3@fe.edu.vn (Luong Vuong Nguyen), kclee@seoultech.ac.kr (Kyungchun Lee), lhpark@seoultech.ac.kr (Laihyuk Park), wsna@kongju.ac.kr (Woongsoo Na), srcho@cau.ac.kr (Sungrae Cho)

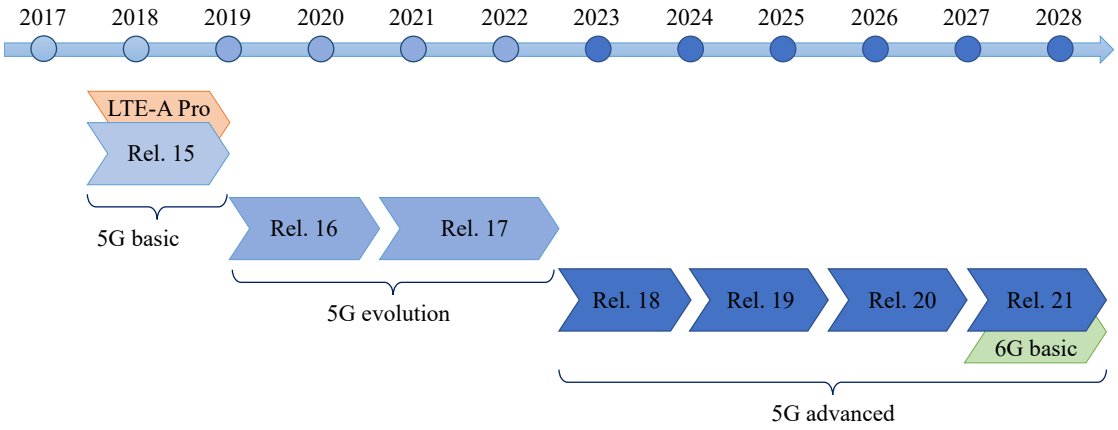


Figure 1: The 3GPP standard milestone for 5G development.

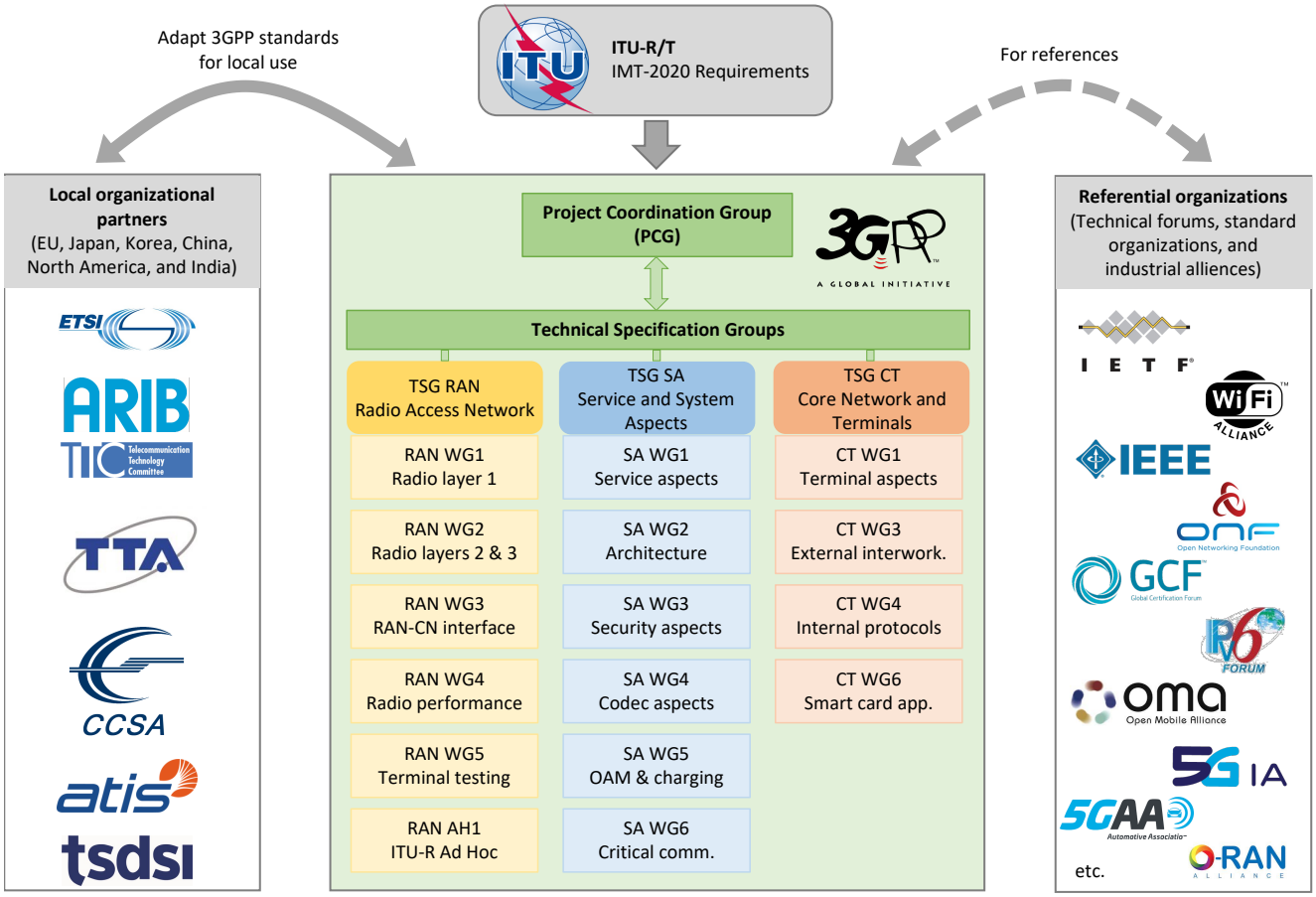


Figure 2: The 3GPP organization structure and its collaboration with partners.

ducing the novel New Radio (NR) architecture, which is ready for implementation in non-standalone networks. This approach enabled 5G infrastructure to be fully compatible with the existing infrastructure in previous generations. Subsequently, 3GPP Rel 16 (in 2020) [4] and Rel 17 (in 2022) [5] specified the requirements and technologies for the next 5G stage (a.k.a. the 5G evolution). In this stage, a fully 5G network was introduced using the NR architecture at the access and a 5G standalone network at the core. The prime case studies and their enablers were standardized for 5G service popularity.

In addition, Rel 18 started to improve the system and service performance using artificial intelligence (AI). The last stage of 5G, 5G Advanced, was planned to be developed from Rel 18 to Rel 21 for the maturation of network intelligence. At Rel 21, there will be a transition from 5G Advanced to the first stage of 6G. Fig. 1 illustrates the complete 3GPP standard milestone planned for 5G development up to 2028.

The 3GPP organizes a management consortium of seven primary members (i.e., local organizational partners from the European Union, Japan, Korea, China, North America, and In-

Table 1: Key abbreviations.

Abbre.	Description
3GPP	3rd Generation Partnership Project organization
5G	Fifth-generation
5GC	Fifth-generation core
ACK	Acknowledgment
AI	Artificial intelligence
ADC	Analog-to-digital
AoA	Angle of arrival
AoD	Angle of departure
BS	Base station
BWP	Bandwidth part
CA	Carrier aggregation
CC	Component carrier
CDMA	Code division multiple access
CP	Control plane
CQI	Channel quality indicator
CSI	Channel state information
CT	Core network and terminals
CU	Central unit
DCI	Downlink control information
DL	Downlink
DM-RS	Demodulation Reference Signal
DRX	Discontinuous reception
DSS	Dynamic spectrum sharing
DU	Distributed unit
EAS	Edge application server
ECS	Edge configuration server
EEC	Edge enabler client
EES	Edge enabler server
EH	Energy harvesting
eMBB	Enhanced mobile broadband
eMTC	Enhanced machine-type communication
eNB	Evolved node B
EPC	Evolved packet core
FAB	Fixed access backhaul
FBE	Frame-based equipment
FDD	Frequency division duplex
FDMA	Frequency division multiple access
FR	Frequency range
GEO	Geosynchronous Earth orbit
gNB	Next-generation node B
GNSS	Global navigation satellite system
HAP	High altitude platform
HARQ	Hybrid automatic repeat request
IAB	Integrated access and backhaul
IDMA	Interleave-division multiple access
IGMA	Interleave-grid multiple access
IoT	Internet-of-Things
ISaC	Integrated sensing and communication
ITU-R	International Telecommunication Union
KPI	Key performance indicator
LBE	Load-based equipment
LBT	Listen-before-talk
LCRS	Low code rate spreading
LEO	Low Earth orbit
LMF	Location management function
LSSA	Low code rate and signature-based shared access
LTE	Long-term evolution
LTE-A	Long-term evolution-advanced
MAC	Medium access control
MBS	Multicast-broadcast service
MEO	Medium Earth Orbit
mMTC	Massive machine-type communication
mmWave	Millimeter wave
MT	Mobile termination
MIMO	Multiple-input multiple-output
ML	Machine learning
NOMA	Non-orthogonal multiple access
MUD	Multiuser detection
MUSA	Multiuser sharing access
MUST	Multi-user superposition transmission
NACK	Negative acknowledgment
NB-IoT	Narrow-band Internet of Things
NCMA	Non-orthogonal coded multiple access
NOCA	Non-orthogonal coded access
NG	New generation
NR	New radio
NR-U	New radio for unlicensed spectrum
NTN	Non-terrestrial network
OFDM	Orthogonal frequency division multiplexing
OFDMA	Orthogonal frequency division multiple access
OMA	Orthogonal multiple access
O-RAN	Open radio access network
OTA	Over-the-air
PCell	Primary cell
PCG	Project coordination group
PDCCH	Physical downlink control channel
PDCP	Packet data convergence protocol
PD-NOMA	Power-domain non-orthogonal multiple access
PDSCH	Physical downlink shared channel
PDU	Protocol data unit
PHY	Physical layer
PMI	Precoding matrix indicator
PRACH	Physical random-access channel
PRS	Positioning reference signal
PUCCH	Physical uplink control channel
PSM	Power saving mode
PSSCH	Physical sidelink shared channel
PTM	Point-to-multipoint
PTP	Point-to-point
PUSCH	Physical uplink shared channel
QoS	Quality-of-service
RACH	Random-access channel
RAN	Radio access network
RDMA	Repetition-division multiple access
RedCap	Reduced capability new radio devices
Rel	Release
RF	Radio frequency
RI	Rank indicator
RLC	Radio link control
RRC	Radio resource control
RRM	radio resource management
RTT	Round trip time
RS	Reference signal
RSMA	Resource-spread multiple access
RSRP	Reference signal received power
SCell	Secondary cell
SCI	Sidelink control information
SCMA	Sparse-coded multiple access
SCS	Subcarrier spacing
SDAP	Service data adaptation
SDT	Small data transmission
SINR	Signal-to-interference-plus-noise ratio
SL	Sidelink
SLA	Service-level agreement
S-NSSAI	Single network-slice selection assistance information
SPC	Short-packet communication
SRS	Sounding reference signal
TB	Transport block
TDD	Time division duplex
TDMA	Time division multiple access
TDoA	Time difference of arrival
TRP	Transmission reception point
TSG	Technical specification group
UAV	Unmanned aerial vehicle
UDP	User datagram protocol
UE	User equipment
UL	Uplink
UP	User plane
URLLC	Ultra-reliable low-latency communication
V2V	Vehicle-to-vehicle
V2X	Vehicle-to-everything
WG	Working group
WSN	Wireless sensor network
WUS	Wake-up signal

dia) and numerous associate members (i.e., referential organizations, including technical forums, standards organizations, and industrial alliances) to certify commercial mobile networks regarding satisfying the IMT-2020 requirements issued by ITU-R for 5G networks, devices, and services. The primary members contribute technical reports and descriptions to maintain specifications and recommendations for 3GPP standards, and associate members refer to the 3GPP standards for their interoperability. These relationships are presented in Fig. 2. To develop an internal structure, 3GPP assigns a project coordination group as the highest decision-making body to coordinate three technical specification group work items, such as the radio access networks (RANs) of six working groups, the service and system aspects of six working groups, and the core network and terminals of four working groups.

A systematic view of the 5G developments through recent 3GPP standard analyses is essential for interested engineers and scholars working in the field. There have been several surveys reviewing different technical aspects of 3GPP standards. For instance, Giordani et al. [6] explored measurement techniques for beam and mobility management, focusing on the design of accurate control schemes tailored for 3GPP NR cellular networks. This tutorial highlighted existing studies on precise beam alignment and considerations for optimal strategy based on deployment environments and system parameters. Jun et al. [7] and Le et al. [8] detailed 3GPP standardization activities for ultra-low latency services. The scope of this review covers the adoption of mobile edge computing (MEC) and time-sensitive communication, aiming to achieve sub-10 ms end-to-end latency in the edge network. Concerning the security perspective, Cao et al. [9] thoroughly discussed security features, requirements, vulnerabilities, existing solutions, and open research issues in 3GPP 5G networks. Regarding IoT services, Jiang et al. [10] extensively investigated the 3GPP Industrial Internet of Things (IIoT) model, comparing it with other indoor models. The survey covered IIoT scenarios, frequency bands, and channel parameters, including path loss and line-of-sight probability, with considerations for antenna height and clutter density. Simi-

▽ **1. Introduction**
▽ **2. Release 15**
- New Radio
- High-Frequency Bands and Radio Frame Structure
- Massive MIMO
- Ultra-Reliable Low-Latency Communication

▽ **3. Release 16**
- NR Non-Orthogonal Multiple Access
- Integrated Access and Backhaul Networks
- NR Positioning
- Latency Reduction
- User Equipment Power Savings
- NR Vehicle-to-Everything Communications
- NR for Unlicensed Spectrum

▽ **4. Release 17**
- RedCap NR Devices
- Non-Terrestrial Network
- NR Multicast-Broadcast Services
- Edge Computing
- Radio Access Network Slicing

▽ **5. Release 18 (as expected)**
- RAN Intelligence
- Network Energy Savings
- Small Data Transmission
- Unmanned Aerial Vehicles
- Low-Power Wake-Up Signal and Receiver
- Dynamic Spectrum Sharing

▽ **6. Release 19 (as planned) and Beyond**
- AI/ML Enhancement for NG RAN
- Integrated Sensing and Communication
- Ambient IoT
- NTN Evolution

▽ **7. Concluding Remarks**

Figure 3: Paper structure and a summary of technologies per release.

larly, Lin et al. [11] and Ali et al. [12] respectively reviewed 3GPP non-terrestrial and V2X networking models, which summarized technical specifications, system designs, and service applications in 5G networks.

These aforementioned studies have significantly contributed to providing an overview of 3GPP activities from various aspects. However, it lacks a comprehensive review of new technologies recommended in the 3GPP release series for 5G access infrastructure and beyond. This observation motivated us to conduct such a survey. In particular, each access network recommendation from Rel 15 to Rel 19 is investigated thoroughly in Sections 2 to 6, respectively. We describe the major technologies and case studies and provide a lesson-learned discussion at the end of each section. For ease of reference, Table 1 provides a list of key abbreviations used throughout the article. In addition, Fig. 3 illustrates the paper outline.

2. Release 15

In June 2018, 3GPP completed the Rel 15 study on technologies for the enhanced long-term evolution (LTE)/LTE-A network, 5G core (5GC) network, and 5G RAN, which defines specifications for the initial 5G phase [3]. This section is dedicated to reviewing the key technological features for the 5G RAN specified in 3GPP Rel 15, including NR, eMBB functionality with the high spectrum and massive multiple-input multiple-output (MIMO) implementation, and URLLC. Technical extensions to LTE/LTE-A systems for mMTC are also implemented in 3GPP Rel 15; however, we do not review them because this section focuses on 5G NR technology.

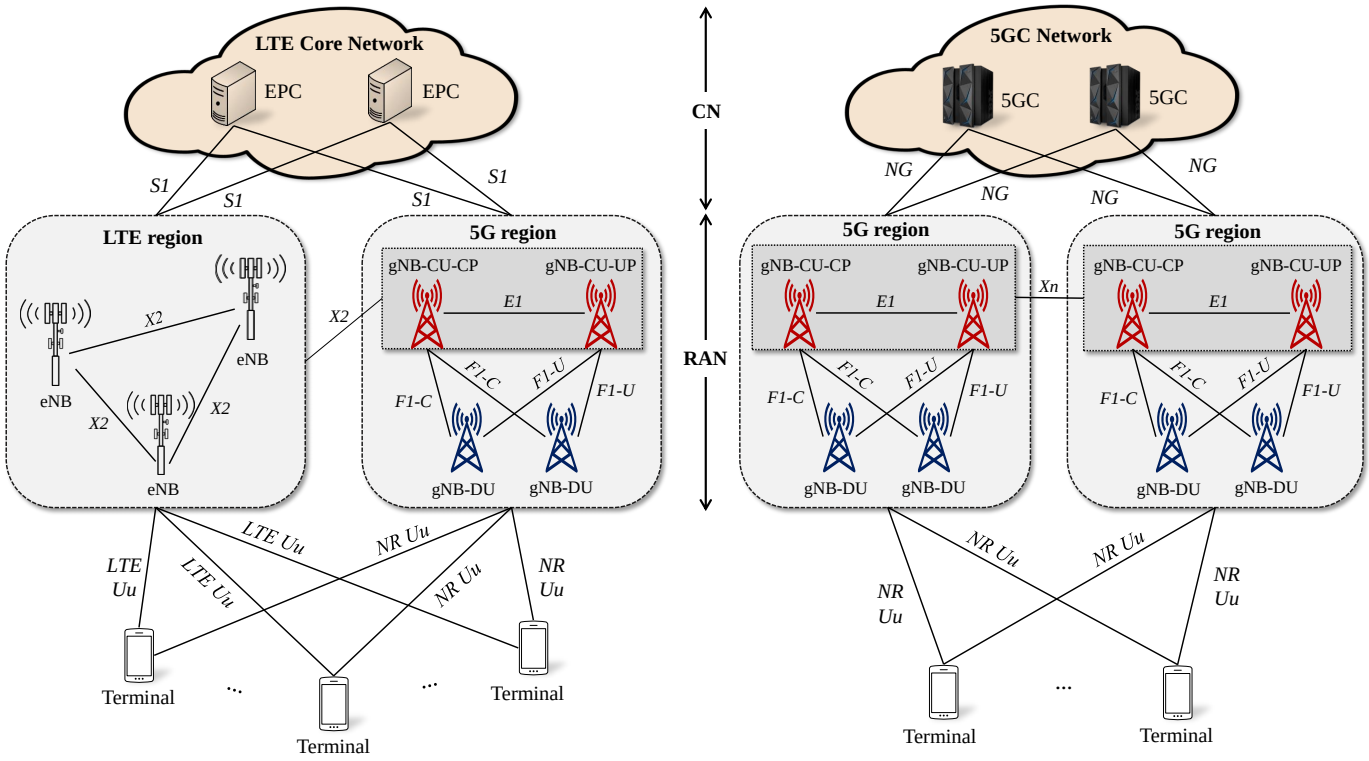
2.1. New Radio

The main evolution from LTE/LTE-A to 5G is the NR technology that satisfies 5G requirements without disrupting the existing LTE and LTE-A systems. One important aspect is that the LTE/LTE-A RAN employs the *LTE Uu* interface to support wireless connectivity for terminals from the LTE evolved Node B (eNB). In contrast, the 5G RAN applies the next-generation (NG) Node B (gNB) functionality with the *NR Uu* interface for 5G terminal connectivity. Additionally, 3GPP Rel 15 provides specifications for the non-standalone and standalone operations for 5G NR, which are described as follows.

- The non-standalone architecture, a.k.a LTE-5G dual connectivity, offers a combination of 5G NR and the existing LTE/LTE-A networks. As depicted in Fig. 4(a), the LTE and 5G regions are connected to the evolved packet core in the LTE core network via the *S1* interface. Moreover, the *X2* interface is leveraged for (*i*) the connections between conventional LTE eNB stations in the LTE region and (*ii*) the connections between LTE eNB and 5G gNB-central unit (CU). This configuration can be a stepping stone toward a full 5G deployment. In such a scenario, LTE services are primarily supported through the *LTE Uu* interface, but they still partially receive 5G NR support through the *NR Uu* interface (e.g., URLLC and eMBB).

- The standalone architecture enables 5G NR systems to operate independently, which excludes the LTE/LTE-A systems. As illustrated in Fig. 4(b), 5G regions are linked to the 5GC in the 5GC network through the *NG* interface. The connections between a set of gNB-CU nodes are conducted via the *Xn* interface. Most importantly, this configuration can be considered the full 5G deployment, where the 5G region offers 5G services to terminals via the *NR Uu* interface only.

For system enhancements, 3GPP Rel 15 has additionally considered the centralized-distributed fashion and functional split for the 5G RAN infrastructure, described as follows. In the centralized-distributed fashion, gNB can be split into gNB-CU and gNB-distributed unit (DU) to enhance scalability, user experience, and DU offloading efficiency. Notably, CU-DU



(a) 5G NR and LTE non-standalone operation

(b) 5G NR standalone operation

Figure 4: Two typical configurations for the 5G NR architecture.

split options for NR RAN logical architectures have been introduced in Rel 14 to facilitate diverse demands [13], including radio resource control (RRC) – packet data convergence protocol (PDCP) (*Option 1*), PDCP – radio link control (RLC) split (*Option 2*), intra-RLC split (*Option 3*), RLC–medium access control (MAC) split (*Option 4*), intra-MAC split (*Option 5*), MAC–physical layer (PHY) split (*Option 6*), intra-PHY split (*Option 7*), and PHY–radio frequency (RF) split (*Option 8*). The detailed description and literature reviews for each split option can be found in [14]. Among the eight CU-DU split options introduced in Rel 14, after extensive discussions, 3GPP defined only two CU-DU split architectures for 5G NR, in addition to the traditional monolithic one, as follows.

- The high-level split with *Option 2* is predominantly recommended for 5G NR because it offers the most straightforward option to standardize while guaranteeing comparable performance [3]. In *Option 2*, gNB-CU is responsible for the service data adaptation protocol (SDAP), RRC, and PDCP radio protocols, whereas the remaining radio protocols (i.e., PHY, MAC, and RLC) reside in gNB-DU. Notably, the implementation of this split option requires coordinating security context parameters when deployed in a topology with different PDCP instances.

- The low-level split with *Option 7* is expected to employ a simple gNB-DU, reduce fronthaul load consumption, and enhance centralization of processing functions for 5G NR [15]. Furthermore, this allows leveraging the benefits of uplink (UL) and downlink (DL) independently. In the UL, gNB-DU hosts the fast Fourier transform and control plane (CP) removal functions, while gNB-CU is responsible for the remaining PHY functions. In the DL, gNB-DU hosts the inverse fast Fourier transform and CP addition functions, while the remaining PHY functions reside in gNB-CU. It is worth mentioning that this split option may require subframe-level timing interactions between a part of PHY in gNB-CU and a part of PHY in gNB-DU. Consequently, gNB-DU cannot solve all synchronization issues.

Within the centralized-distributed fashion, each gNB-DU is managed by only one gNB-CU, whereas the gNB-CU leads the networking functionality for one or more gNB-DU nodes through the *F1* interface [16]. This is equivalent to the fact that the CU-DU split can increase the number of cells managed by each gNB-CU, allowing for more handovers to be handled through intra-gNB mobility. This approach has a significantly smaller impact compared to handovers handled by inter-gNB mobility when the anchor point of the devices remains constant. As a result, the CU-DU split can maximize the ratio of intra- versus inter-gNB handovers. Furthermore, when terminals may connect to multiple different gNBs, only one of gNBs (the anchor gNB) is responsible for processing split data streams via PDCP, which can cause load imbalance and inefficient resource usage among gNBs. Motivated by this drawback, the CU-DU split enables the offloading of PDCP aggregation to the gNB-

CU, where resource sharing or pooling can efficiently handle the task.

Furthermore, the functionality of gNB-CU can be split into a gNB-CU-CP and gNB-CU-user plane (UP), where the interconnection between two functional plane parts of gNB-CU (i.e., gNB-CU-CP and gNB-CU-UP) are specified by the *E1* interface [17]. The gNB-CU-CP and gNB-CU-UP are connected to the gNB-DU through the *F1-C* and *F1-U* interfaces, respectively. This approach enhances the functional possibility of the 5G RAN infrastructure in at least four aspects: (*i*) optimizing the location of various RAN functions in a centralized-distributed fashion, (*ii*) efficiently supporting the radio resource isolation for network slicing, (*iii*) independently scaling the CP and UP capacity, and (*iv*) addressing the bandwidth scarcity problem for the transport between gNB-CU and gNB-DU, especially in the context of the high spectrum, massive MIMO, and URLLC.

From the enhanced functionality 5G NR perspective, some typical academic studies using 3GPP Rel 15 can be reviewed below. In [18], a fast data retransmission mechanism has been introduced to assure the data delivery between CU and DU in both the non-standalone and standalone architectures of 5G NR using the CU-DU split *Option 2*, where the retransmitted data packets can be identified and handled with high priority.

Subsequently, Xu *et al.* [19] proposed an improved solution to achieve fast-centralized retransmission of lost PDCP protocol data units (PDUs) in the intra-CU inter-DU handover scenario of the CU-DU split 5G NR. The system-level simulation results in [19] have demonstrated that the CU-DU split using *Option 2* offers a better user experience than that of using *Option 3*.

In [20], a field-programmable gate array-based optimization of a gNB-DU receiver was designed to optimize the power, throughput, maximum operation frequency, and latency of the 5G NR DU utilizing the CU-DU split *Option 7.1*. The simulation results have revealed a direct dependence on the data type, thereby making information available for a design decision. The modulation scheme is almost agnostic, providing reliable information for an optimized gNB-DU implementation.

Furthermore, the work [21] proposed a novel multilink mechanism, namely a single grant-based UL, for the 5G NR CU-CP/CU-UP/DU split architecture. The proposed scheme can be briefly described by three main steps: (*i*) gNB-CU-CP is responsible for configuring UL grants for a given UE and provides this UL grant information to all gNB-DUs; (*ii*) each gNB-DU recognizes to independently receive and process a single UL signal received from the anchor UE; (*iii*) gNB-CU-UP processes the desired signal by using sequence numbering-based duplicate removal. The system-level simulation results have revealed that the proposed framework significantly improves the throughput, latency, and reliability compared to a single link mechanism.

It is worth mentioning that the O-RAN Alliance organization has extended the functional split *Option 7.2* of 3GPP 5G NR for increased disaggregation, which is referred to as open RAN (O-RAN) [22]. O-RAN disaggregates gNB functionalities into a CU, a DU, and a radio unit. Additionally, O-RAN connects these split elements to a near-real-time RAN intelligent controller with a configuration period of 10 – 1000 ms and a non-real-time RAN intelligent controller with a configuration period exceeding 1 s. When O-RAN does not fall within the scope of this work, we do not delve into detailed reviews. For comprehensive information and literature reviews on O-RAN, please refer to [22].

The enhanced functionalities for 5G NR compared to LTE/LTE-A in 3GPP Rel 15 are to realize stringent URLLC satisfaction and enable eMBB with high-frequency bands and massive MIMO deployment. Subsequently, the NR-based functionalities for high spectrum and bandwidth, massive MIMO, and URLLC in 3GPP Rel 15 are reviewed in the subsequent subsections.

2.2. High-Frequency Bands and Radio Frame Structure

The significant differences between 5G NR and LTE/LTE-A are the characteristics of extremely high-frequency bands and their large channel bandwidth. The existing LTE/LTE-A systems have provided services using 2 GHz and 800 MHz frequency bands, where the maximum carrier bandwidth configured for the user equipment (UE) is 20 MHz [23]. In contrast, 5G NR in 3GPP Rel 15 is expected to leverage the high spectrum with broader bandwidth of the sub-6 GHz (or sub-7 GHz) and millimeter wave (mmWave) frequency bands to adopt various deployment scenarios, which are categorized into frequency range (FR) 1 and 2, respectively. The applicable FR, newly defined NR bands, supported channel bandwidth, and defined parameters for multiple numerologies (i.e., subcarrier spacing (SCS), the cyclic prefix, number of subframes per radio frame, number of slots per subframe, and number of orthogonal frequency division multiplexing (OFDM) symbols per slot corresponding to SCS) for each FR classification in 3GPP Rel 15 are summarized in Table 2.

As observed in Table 2, FR1 specifies the applicable FR below 7 GHz (i.e., 450 MHz – 7.125 GHz). There are relatively few technical issues because the use of FR1 is similar to that of LTE/LTE-A systems (i.e., the LTE refarming band) in which the technological aspects have matured. One disadvantage is that an orderly wide frequency band cannot be guaranteed because most frequencies are already occupied in use. Motivated by this, n77 (3.3 – 4.2 GHz), n78 (3.3 – 3.8 GHz), and n79 (4.4 – 5.0 GHz) are new frequency bands defined for 5G NR in FR1 on a first-come, first-served basis. Moreover, FR2 specifies the mmWave band (i.e., 24.25 – 52.6 GHz), where an orderly wide frequency band can be guaranteed due to the hardly used nature of this band. It prevails in gaining easy support for high speed, a large capacity, and data transmissions but suffers from the large over-the-air (OTA) signal attenuation, motivating the technical discussion for the massive MIMO application in Section 2.3. It is worth noting that only the OTA testing methodology is specified in FR2, whereas both the conducted testing and OTA testing methodologies can be used in FR1. For the conducted testing method, antenna connectors are still accessible. However, an OTA testing method does not prefer antenna connectors due to the excessive complexity, high cost,

Table 2: Summary of the newly defined NR bands, channel bandwidth, SCS, and multiple numerologies for FR1 and FR2 in 3GPP Rel 15.

Classification	Applicable frequency band	Newly defined NR bands	Channel bandwidth (MHz)	SCS (KHz)	Cyclic prefix	Subframes per radio frame	Slot(s) per subframe N_{slot}	OFDM symbols per slot N_{symbol}	Supported channels	
									Data	Sync.
FR1	450 MHz – 7.125 GHz	n77 (3.3 – 4.2 GHz) n78 (3.3 – 3.8 GHz) n79 (4.4 – 5.0 GHz)	5, 10, 15, 20, 25, 30, 40, 50, 60, 80, 90, 100	15	normal	10	1	14	Yes	Yes
				30	normal		2	14	Yes	Yes
				60	normal		4	14	Yes	No
					extended			12		
FR2	24.25 GHz – 52.6 GHz	n257 (26.5 – 29.5 GHz) n258 (24.25 – 27.5 GHz) n260 (37.0 – 40.0 GHz) n261 (27.5 – 28.35 GHz)	50, 100, 200, 400	60	normal	10	4	14	Yes	No
					extended			12		
				120	normal		8	14	Yes	Yes
				240	normal		16	14	No	Yes

and physical size limitations, especially in massive MIMO systems operating in FR2. In FR2, n257 (26.5 – 29.5 GHz), n258 (24.25 – 27.5 GHz), n260 (37.0 – 40.0 GHz), and n261 (27.5 – 28.35 GHz) are the new frequency bands that have not been supported for LTE/LTE-A in the literature.

In summary, 5G technologies with a sub-6 GHz frequency band can cover a larger geographical area for 5G coverage with their mature techniques, but they cannot provide a significantly higher speed DL compared with that of LTE/LTE-A. Moreover, mmWave 5G technologies can provide class-leading DL speeds with low latency but suffer from less geographical area coverage.

Unlike LTE/LTE-A systems with 100 MHz of the maximum aggregated bandwidth, the maximum channel bandwidth for a single component carrier (CC) in 5G NR is 100 MHz for FR1 and up to 400 MHz for FR2, respectively. Consequently, with the carrier aggregation (CA) support possibility of up to 16 CCs, the maximum aggregated bandwidth in 5G NR is 1.6 GHz for FR1 and 6.4 GHz for FR2, respectively. Furthermore, 5G NR supports multiple OFDM SCS for data transmission and/or synchronization, where defined parameters for flexible frame structures with different numerologies in FR1 and FR2 are provided in Table 2. The NR frame structure is designed appropriately according to the different choices for SCS. An illustration of the radio frame structure in 5G NR is presented in Fig. 5, where N_{slot} and N_{symbol} denote the number of slots per subframe and the number of OFDM symbols per slot, respectively. Furthermore, an extended cyclic prefix is additionally defined for the 60 KHz SCS, in which there are only 12 OFDM symbols per slot. Notably, not every numerology can be applied to all data and control channels. Specifically, SCS of 60 KHz only supports data channels, such as the physical DL shared channel (PDSCH) and physical UL shared channel (PUSCH). Contradictorily, SCS of 240 KHz is only used for synchronization channels, including the primary synchronization signal (PSS), secondary synchronization signal (SSS), and physical broadcast channel (PBCH). It is worth mentioning that, in addition to the SCSs introduced in Table 2, Rel 17 further incorporated SCSs of 480 KHz and 960 KHz to enhance the supported channels of both data transmission and synchronization [24].

We are currently witnessing a global rush to 5G deployment. During the initial phase of 5G devoted to 3GPP Rel 15, three NR new bands in FR1 and four NR new bands in FR2 expose the essential investigation for the spectrum allocation plans in each country worldwide. While technical standards for 5G services were still being finalized at this stage, the European Union (EU), United States of America (USA), Korea, Japan, and China were still planning to be the first leaders to deploy working commercial 5G networks. Fig. 6 illustrates the spectrum allocation plans for 5G NR to these leading countries, where the data was collected from the following materials [25, 26, 27, 28]. It is worth noted that upper mid bands, such as 7.125 − 24.25 GHz, were triggered at a later stage in Rel 16. This frequency range is referred to as FR3 to differentiate it from FR1 and FR2. Subsequently, Rel 17 defined 5G NR in the frequency range of 52.6 − 114.25 GHz, called FR4. To observe the spectrum regulations of RF3 and RF4, please refer to [29] and [30], respectively. As of Rel 18, the frequency range greater than 114.25 GHz is expected to be explored for enhanced applications of 5G Advanced [31].

In academic studies, Coll-Perales *et al.* [32] proposed a sub-6 GHz assisted mmWave MAC that decouples the mmWave data plane using the 802.11ad standard and CP using the 802.11p standard. The proposed framework allows offloading mmWave MAC control functions (e.g., beamforming, link availability identification, and scheduling) to a sub-6 GHz vehicle-to-everything (V2X) technology standardized by 3GPP Rel 15. The numerical results have shown reduced control overhead, delay degradation, and improved spatial sharing, network capacity, and scalability.

Subsequently, [33] introduced a novel compact-size dual-function antenna that operates simultaneously at 3.5 GHz and 28 GHz for 5G NR, where a frequency reconfigurability technique was adopted. To enhance diversity and multiplexing performance, a 8×8 MIMO prototype was considered. The numerical results have shown good agreement with satisfactory MIMO performance and safety guidelines.

Sang *et al.* [34] proposed a novel dual-band planar antenna array by using recombining technology for 5G NR. A 4×4 antenna array scale is designed for the sub-6 GHz bands (i.e., 4.5 – 6 GHz), whereas a 4×8 prototype is considered for operation in mmWave bands (i.e., 24.5 – 26 GHz). OTA experimental results have verified the compactness and good performance of the proposed prototype.

Furthermore, a low-pass filter-based integrated 5G smartphone antenna was proposed in [35] for multi-band operation

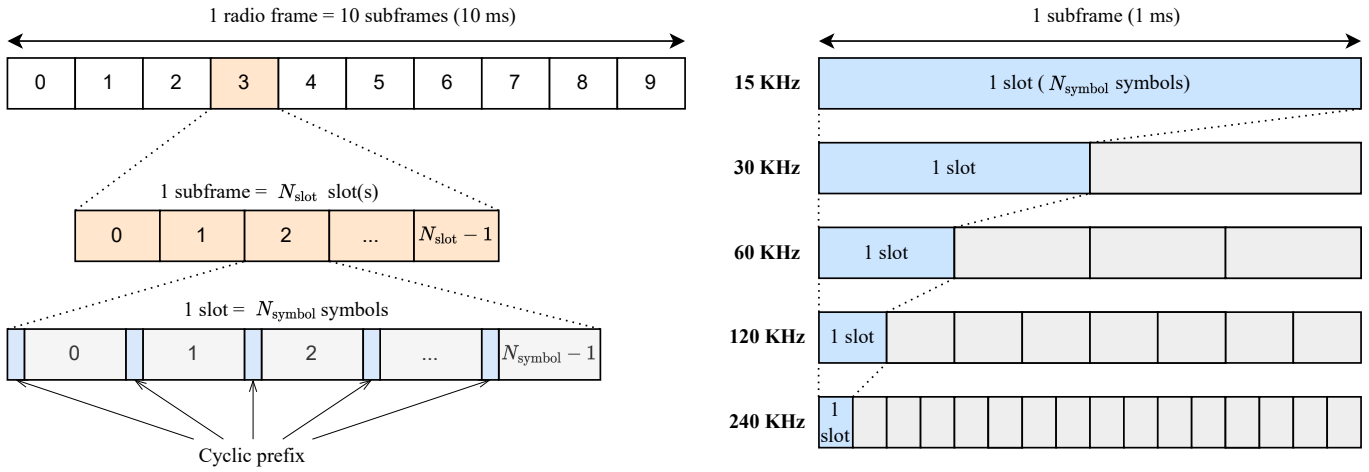


Figure 5: 5G NR frame structure defined in 3GPP Rel 15.

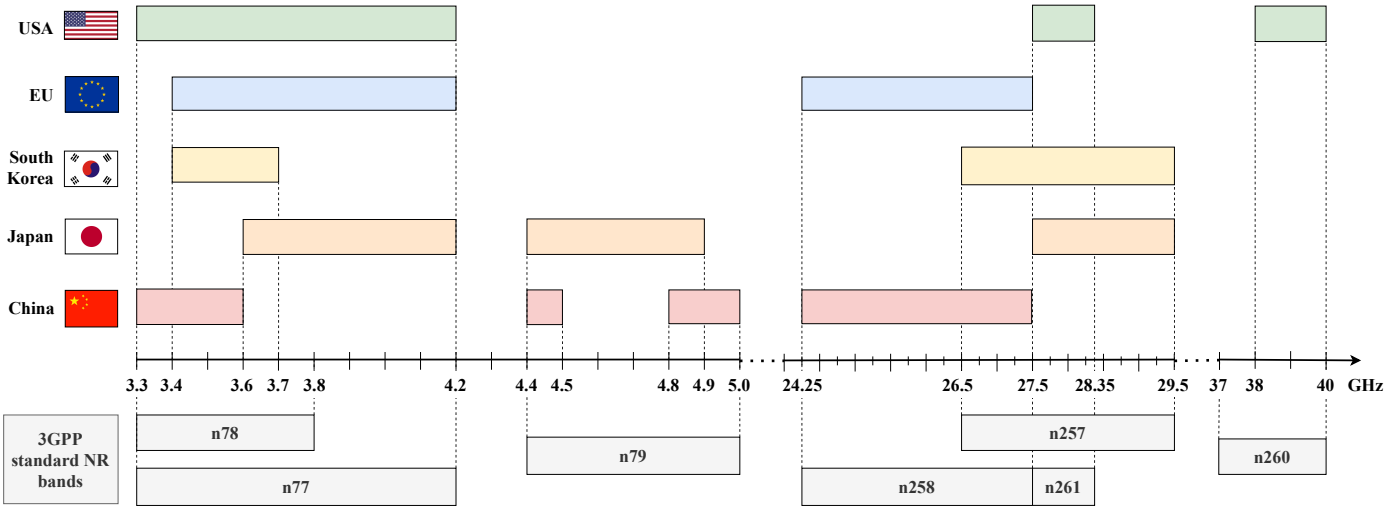


Figure 6: Spectrum allocation plans for 5G NR specified by 3GPP Rel 15 [25, 26, 27, 28].

at 3.5 GHz, 28 GHz, and 38 GHz, where a fabricated 8×8 array scale prototype was designed. The numerical results have validated satisfactory radiation performance and ensured safety across all considered operating bands.

In [36], multi-frequency cooperative communication at 3.3 GHz (FR1), 6.5 GHz, 15 GHz, and 28 GHz (FR2) was investigated to optimize spectrum utilization and enhance system performance in urban micro and outdoor-to-indoor environments. The simulation results have elucidated channel characteristics such as path loss, shadowing, frequency-dependent clustering, delay spread, Rician factor, and correlation properties. Additionally, it underscored performance tradeoffs regarding cell coverage radius, data rate, and bit error rate among the considered frequency bands.

2.3. Massive MIMO

Massive MIMO is one of the key components of 5G NR technologies to achieve extra degrees of freedom for increasing the system throughput and substantial beamforming gains for coverage improvement, where the support of up to 256 antenna elements at each 5G gNB and up to 32 antenna elements at each UE terminal is specified in 3GPP Rel 15 [3]. In practice, numerous antenna elements can be assembled into multiple panels aiming for cost and power reduction. The 5G NR massive MIMO technology also provides the capability of multilayer data transmission through multiple antenna ports.

Notably, the antenna port is a novel concept introduced in 5G NR for massive MIMO. Before delving into its definition and illustration, we first introduce new reference signals (RSs) as preliminaries. The RSs are predefined signals occupying specific resource elements within the DL/UL time-frequency grid. The 5G NR specification includes several types of RSs transmitted differently for different purposes by a terminal. The functional descriptions of RSs, i.e., demodulation RS (DM-RS), tracking RS, phase tracking RS, channel state information (CSI) RS, and sounding RS (SRS), are briefly summarized in Table 3. Among them, CSI-RS is a key component for beam management and the relevance of antenna ports in 5G NR. Notably,

Table 3: Summary of RSs and beam-functional elements for 5G NR beam management.

Classification		Functional description
Reference signals	DM-RS	For the PDSCH, the DM-RS is used for DL channel estimation as part of coherent demodulation, envisioned in the resource blocks for DL transmission. For the PUSCH, the DM-RS allows the 5G gNB to demodulate the UL transmission coherently.
	Tracking RS	The tracking RS is a sparse RS that supports terminals in time-/frequency-domain tracking.
	Phase tracking RS	The phase tracking RS, also considered an extended version of DM-RS, is used for phase-noise compensation that enables terminals to track the phase and mitigate performance loss for PDSCH and PUSCH.
	CSI-RS	The CSI-RS, a DL RS, is transmitted by 5G gNB for UE to use in order to estimate DL CSI. Additionally, the CSI-RS is also leveraged for mobility measurement, gNB transmission beamforming measurement, and frequency/time tracking.
	SRS	The SRS, an UL RS, is transmitted by terminals used for UL channel estimation at 5G gNBs.
Beam management	Beam determination	Each beam is determined by a source RS and indicated by its corresponding index, facilitating the 5G gNB and terminals to select suitable transmit and receive beams for the subsequent transmission of DL and UL channels.
	Beam measurement and reporting	The terminal measures beam quality and reports it to the 5G RAN. The network leverages this information when determining appropriate beams to use and configure when communicating with the terminal.
	Beam sweeping and refinement	These procedures enable the 5G gNB to sweep and refine beams (i.e., using narrower beams) and track beams when the terminal moves around or changes orientation.
	Beam association for DL and UL channels	This functionality associates the RSs of DL and UL channels with a beam's source RS. The association is based on quasi-colocation and spatial transmitter parameters.
	Beam failure recovery	This procedure supports a rapid realignment of the gNB and terminal beams whenever a dominated beam is lost due to a sudden blockage or the fast movement/rotation of the terminal.

Table 4: Supported configurations of the virtual logical antenna array specified by 3GPP Rel 15.

CSI-RS port	Virtual logical antenna port array configuration	
	Single panel (N_h, N_v)	**Multi-panel** (N_p, N_h, N_v)
$P = 2$	(1,1)	-
$P = 4$	(2,1)	-
$P = 8$	(2,2) or (4,1)	(2,2,1)
$P = 12$	(3,2) or (6,1)	-
$P = 16$	(4,2) or (8,1)	(2,4,1), (4,2,1), or (2,2,2)
$P = 24$	(4,3), (6,2), or (12,1)	-
$P = 32$	(4,4), (8,2), or (16,1)	(2,8,1), (4,4,1), (2,4,2), or (4,2,2)

Table 5: 5G NR antenna port series specified by 3GPP Rel 15.

Antenna port	DL	UL
0-series	-	DM-RS for PUSCH
1000-series	PDSCH	SRS, PUSCH
2000-series	Physical DL control channel (PDCCH)	Physical UL control channel (PUCCH)
3000-series	CSI-RS	-
4000-series	PBCH or synchronization signal block	Physical random-access channel (PRACH)

CSI-RS in 5G NR can be seen as a counterpart of cell RS in LTE. Specifically, cell RS is a mandatory signal transmitted in LTE, whereas CSI-RS can be configured by RRC to be transmitted in 5G NR, which is not a mandatory signal. There are several options for the number of CSI-RS ports, denoted by P, including $P \in \{1, 2, 4, 8, 12, 16, 24, 32\}$. Here, $P = 1$ is primarily utilized for the tracking RS, while $P \geq 2$ is used for CSI measurement [37]. Furthermore, CSI-RS can be flexibly transmitted in any OFDM symbols and subcarriers, as configured in the RRC message.

In addition to the definition of CSI-RS ports, it is worth mentioning that the number of layers (a.k.a. the number of data streams) is equivalent to the number of DM-RSs. In 5G NR, the DL transmission supports a maximum number of eight layers for single-user massive MIMO systems, whereas the UL transmission supports a maximum number of four streams for such systems. For multiuser massive MIMO systems, a maximum number of 12 transmission layers is supported for both the DL and UL transmissions.

Unlike conventional physical antennas, an antenna port refers to a virtual logical antenna, representing a specific and unique radio channel. As introduced in [37], an antenna port is strictly defined as "such that the channel over which a symbol on the antenna port is conveyed can be inferred from the channel over which another symbol on the same antenna port is conveyed." From a conceptual standpoint, the receiving side can assume that two transmissions share the same radio channel only when they utilize the same antenna port. It is intuitively understandable that if two signals are transmitted through different physical antennas or spatial filters, they will experience difference channels, resulting in distinct antenna ports. However, even if transmitted through the same physical antennas using the same spatial filter, differing radio channels and antenna ports can occur if the RSs used for channel estimation are different. Therefore, the critical aspect of an antenna port lies in channel estimation obtained through transmitted RSs.

Notably, when CSI-RS is generated in the final stage of PHY for MIMO channel estimation, the high-level concept of an antenna port generally refers to the CSI-RS ports. For each value of P, 3GPP Rel 15 defined various options for the structure of the virtual logical antenna port array [38], as shown in Table 4. Here, N_h and N_v represent the number of cross-polarized virtual antenna ports on the horizontal and vertical directions on a single panel, respectively, while N_p defines the number of supported panels. Due to support of cross-polarized antennas, the number of CSI-RS ports is precisely shaped by $P = 2N_pN_hN_v$. Furthermore, in 5G NR specified by 3GPP Rel 15, various antenna-port series were assigned based on the different phys-

ical channel utilization and its corresponding RS, as summarized in Table 5 [37]. It is worth mentioning that, in addition to the antenna-port series introduced in Table 5, Rel 16 further included the 5000-series for DL positioning RS (PRS) [39].

To ease the understanding of antenna ports incorporated with layers, CSI-RS ports, and a physical antenna array, we illustrate their collaborative operation in Fig. 7. The first stage involves mapping raw data from L layers, denoted by $\mathbf{d} = [d_0, d_1, \ldots, d_{L-1}]^T$, to a CSI-RS port array. This is achieved through the multiplication of $\mathbf{Wd}$, where $\mathbf{W}$ represents the precoding matrix indicator (PMI) matrix (known as the precoder) with dimension $P \times L$. Subsequently, the second stage encompasses the mapping from a CSI-RS array to a virtual logical antenna array with cross-polarized antenna ports. Finally, the last stage illustrates the mapping of antenna ports to a physical antenna array. Notably, the dimension of a physical antenna array should be greater or equivalent to the dimension of a virtual logical antenna array, especially in beamforming applications, where it is advisable for the physical array to be significantly larger. In this example, we assume that four physical antennas share the same antenna port.

It is widely known that beam management defines a set of procedures that enable beam transmission and reception in the desired direction with narrow angular coverage to focus the transmitted energy optimally, leading to a significant antenna gain improvement [40]. Beam management in 5G NR is completed by introducing beam-functional elements that are responsible for specific beam procedures, including beam determination, beam measurement and reporting, beam sweeping and refinement, beam association, and beam failure recovery [41]. The functional descriptions of beam procedures are briefly summarized in Table 3.

For DL transmit beamforming, beam weights are generated based on CSI feedback, known as CSI reporting, from UE, where the CSI feedback includes the rank indicator (RI), PMI, and channel quality indicator (CQI). These weights are used to allocate multiple layers to (*i*) a single terminal in single-user massive MIMO systems or (*ii*) different terminals in multiuser massive MIMO systems. In these cases, the beamforming weight should be carefully chosen to maximize the beam gain toward the desired terminal and minimize the leakage power toward other terminals. Fundamentally, RI represents the number of possible layers for DL under specific channel conditions, which is beneficial for computing PMI and CQI. The PMI and CQI selection processes are summarized as follows.

- The PMI selection is conducted based on the codebook type, the number of transmission layers, and panel dimensions. In Rel 15, the 5G NR DL defined three types of supported codebooks: Type-I single-panel, Type-I multi-panel, and Type-II codebooks [38]. The NR Type-I codebook family has roughly the same construction as the LTE codebook but with more complexity, supporting more diverse types of predefined matrices. The NR Type-II codebook has only one conventional table but with a very complicated formula defining the precoding matrix. Please refer to [38, Tables 5.2.2.2.1-1 to 5.2.2.2.1-12] and [38, Tables 5.2.2.2.2-1 to 5.2.2.2.2-6] for specific parameters of the NR Type-I codebook family, while [38, Tables 5.2.2.2.3-1 to 5.2.2.2.3-5] provides specific parameters for the NR Type-II codebook. The NR Type-II codebook can apply a more sophisticated precoding matrix than the NR Type-I codebook using such a complicated form with many parameters. Thus, Type I only supports a specific single beam, whereas Type II supports a transmit beamforming with a group of beams by linearly combining them within the group. It is worth mentioning that Rel 16 further included the enhanced Type-II codebook [42].

- The CQI selection can be performed based on the presence or absence of the given signal-to-interference-plus-noise ratio (SINR) lookup tables. For CQI selection using lookup tables [38, Tables 5.2.2.1-2 to 5.2.2.1-4], SINR values are mapped to codewords, and CQI is computed based on SINR thresholds in accordance with block error rate requirements, with the option to adjust the CQI index if needed. In contrast, for CQI selection without using lookup tables, CQI is calculated by analyzing SINR values across all layers associated with a reported PMI. The system then selects the appropriate CQI combination based on the number of codewords, following specified block error rate conditions.

For UL transmit beamforming, the terminal transmits multiple SRS symbols to the 5G gNB. Subsequently, the 5G gNB measures them and identifies the best beamforming weight corresponding to that terminal, where the UL 5G NR beam management is based on codebook-based and non-codebook-based schemes [38]. In the codebook-based classification, the 5G gNB measures the UL channel on the antenna ports of the SRS resource and determines the rank and codebook index of the PUSCH. In the non-codebook-based classification, the terminal transmits multiple antenna-precoded SRS resources according to the DL CSI-RS channel measurements. Subsequently, the 5G gNB measures the channel on the SRS resources. The UL precoding at the terminal follows the precoded SRS resources indicated by the 5G gNB.

Unlike the beamforming technology in traditional LTE/LTE-A systems, which relies solely on the digital domain, 5G NR allows additional implementation of analog beamforming [43]. Digital beamforming typically involves manipulating the phase and amplitude of signals in the baseband or digital domain before upconversion to RF for transmission, enabling adaptive beamforming based on channel conditions. This manipulation is usually achieved through sophisticated signal processing techniques such as matrix manipulations. In this scenario, multiple beams (one per each UE) can be formed simultaneously from the same set of antenna elements. Meanwhile, analog beamforming is more commonly associated with beam steering by adjusting the phase and amplitude of signals in the RF domain at the level of individual antenna elements [43]. This predominantly provides enhanced coverage, especially in environments with considerable propagation loss in FR2. However, analog beamforming allows only one beam per

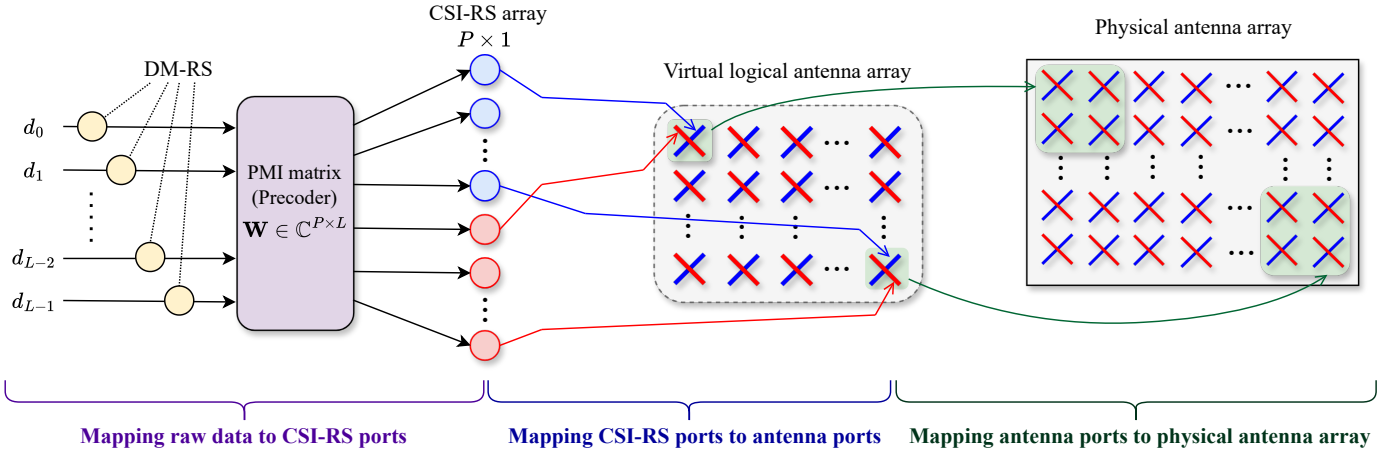


Figure 7: An illustration of the collaborative operation among layers, CSI-RS ports, antenna ports, and a physical antenna array.

set of antenna elements, limiting data stream multiplexing. It is worth mentioning that hybrid beamforming has been proposed in academia with the aim of harnessing the advantages of both beamforming techniques, but it has not yet been standardized within 3GPP releases. In hybrid beamforming, analog beamforming is used for coarse adjustments, while digital beamforming provides fine adjustments and adaptability. For a comprehensive understanding and literature reviews on hybrid beamforming, please refer to [44].

Additionally, 5G NR developed a flexible framework for massive MIMO with the support of both the time-division duplex (TDD) and frequency division duplex (FDD) operation modes. For TDD, CSI could be acquired by exploiting the UL pilot signal through UL and DL channel reciprocity. For FDD, when channel reciprocity between the UL and DL is unavailable, the 5G gNB obtains the CSI by exploiting the DL pilot signal and UL feedback of the terminal.

In academic studies, the applications of 5G NR massive MIMO systems based on the 3GPP Rel 15 specification have been examined [45, 46, 47, 48]. Specifically, a study [45] provided system-level simulation results to demonstrate the performance characteristics of 5G NR massive MIMO systems based on the 3GPP Rel 15 specification (i.e., antenna array modeling with beamforming, UL/DL codebook, TDD/FDD transmission schemes, and single-user/multiuser massive MIMO), where the influence of the array configuration, deployment scenario, and transmission schemes on system performance was investigated. For the same antenna array configurations, the system performance in terms of the spectral efficiency of the 5G NR massive MIMO is significantly better than that of the LTE massive MIMO.

In the study [46], system-level simulations of a 5G NR massive MIMO system with FDD transmission using DL pilot training and UL feedback were conducted based on the 3GPP Rel 15 standard. Besides that, the authors in [46] also proposed a novel channel estimation algorithm, namely compression-based linear minimum mean squared error, associated with a novel adaptive estimate, namely the block sparsity adaptive matching pursuit. The superiority of the proposed methodology in terms of the mean squared error and time complexity over conventional benchmarks was demonstrated.

In addition, in the context of the 5G NR massive MIMO based on the 3GPP Rel 15 specification, Liu *et al.* [47] proposed a phase preprocessing algorithm for the compression process and a variable bit width for compressing phase quantization to efficiently reduce the feedback overhead caused by a DL NR Type-II codebook that consists of a wideband beam group and beam combination coefficients of subbands.

Furthermore, learning algorithms have become a common emergency trend as the world gradually moves toward AI. An extreme learning machine-based receiver was designed for 5G NR multiuser massive MIMO systems [48], where the 5G NR data structure with the TDD mode using NR RSs was specified in 3GPP Rel 15. Extensive link-level radio network simulations were conducted to confirm the practical design of the proposed framework in terms of the sum spectral efficiency, average bit error rate, and time complexity.

2.4. Ultra-Reliable Low-Latency Communication

Up to 3GPP Rel 14, reliability and latency were treated rather independently. 3GPP Rel 15 laid the foundation for joined aspects, i.e., URLLC. In 5G NR systems, latency can be roughly divided into two major factors: CP latency and UP latency [49]. The CP latency refers to the transition time for a terminal to switch from an idle state (i.e., a battery-efficient state, where a terminal is not connected with RRC) to an active state (i.e., the start of a continuous data transfer). In contrast, UP latency (a.k.a. radio interface latency) is measured by the time it takes to deliver an internet protocol packet from the transmitter PDCP to the receiver PDCP via the radio interface in UL and DL communications. For the latency performance KPI specified in 3GPP Rel 15, CP latency is expected to be lower than 10 ms, whereas UP latency should be 0.5 ms for both UL and DL directions. When the application performance primarily depends on UP latency, UP is the crucial focus of interest for low-latency communications. In addition, the KPI of mobility interruption time is almost free of latency (i.e., 0 ms) for intra- and inter-frequency handovers of intra-NR mobility.

From the URLLC perspective, the ultra-reliability KPI can also be associated with a low latency requirement, especially for delay-sensitive information with trusted communications. Reliability is the success probability of transmitting a predefined number of bytes within a specific delay [49]. In accordance with a general URLLC case specified in 3GPP Rel 15, both the ultra-reliability of 0.99999 (five nines) and the low UP latency of 1 ms must be satisfied with a reference payload size of 32 bytes. For other specific Internet of Things (IoT) applications (e.g., factory automation, process automation, smart grids, intelligent transport systems, professional audio, virtual reality, serious gaming, and health care), latency can be relaxed to 10 ms or longer. In contrast, reliability can be guaranteed at six nines or even higher [50, 51, 52, 53].

For instance, the URLLC requirements for vehicle-to-everything communications are five nines of reliability and $3 \sim 10$ ms of UP latency [53]. In the guest editorial special issue [54], the authors recommended 11 outstanding publications that included a comprehensive survey and technical design papers for 5G URLLCs from industry and academia, which were expected to stimulate enormous innovations for 5G URLLC realization and performance improvement from the global research community.

Within the URLLC scope, some typical academic studies using 3GPP Rel 15 specifications can be reviewed below. Ji *et al.* [55] highlighted essential URLLC requirements and presented enabling PHY technologies specified in 3GPP Rel 15, where the advanced studies for packet and frame structure, scheduling schemes, transceiver architecture for dynamic numerology adaptation and simultaneous decoding schemes, and reliability improvement techniques are presented. Additionally, [55] indicated that advanced channel coding schemes with short-packet communication (SPC) should be worth considering, where the repetitive transmission scheme using time-domain resources is feasible because of the short slot-length implementation. In [56], the authors investigated multihop relay networks gained along with the spatial diversity of MIMO implementation under the SPC context to facilitate URLLC requirements, where the individual and joint optimization problems of power allocation and relay location configurations were also studied to achieve the enhanced ultra-reliability and significant power savings. The results demonstrated that SPC is beneficial for satisfying the ultra-reliability to at least five nines and the low-latency constraint of at most 10 ms. Popovski *et al.* [57] discussed the principles of wireless access for URLLCs based on 3GPP Rel 15 specifications, where the communication-theoretic principles of URLLCs were provided on the trade-off relationship between latency, reliability, bandwidth, packet size, and finite-block-length treatment. Frame synchronization, traffic arrival strategies, massive MIMO, multiconnectivity, and the statistical methodology of ultra-reliability guarantees were also investigated. Subsequently, the study [58] described the functional features of LTE-5G multiconnectivity per the URLLC requirements specified in 3GPP Rel 15. Shortened slot structures, wider OFDM SCS, and grant-free UL access were introduced to facilitate the low-latency constraint to comply with this requirement. In contrast, robust coding, robust modulation, diversity schemes, multiconnectivity with packet duplication on PDCP, and CA were the options to enhance the ultra-reliability. In addition, compared to eMBB services, reduced spectral efficiency requires more attention to boosting URLLC services. Esswie *et al.* [59] presented the URLLC outage performance under the optimal interference-free and intercell cross-link interference constraints for 5G NR dynamic-TDD systems with various 5G NR functionalities specified in 3GPP Rel 15. For example, the functionalities include the frame structure, offered sporadic packet arrivals, transmission time interval duration, channel SCS, and proportional-fair and grant-free UL scheduling. Based on the system-level simulation results, the superiority of the flexible-FDD over the dynamic-TDD was also demonstrated.

2.5. Summary and Discussion

Generally, 3GPP Rel 15 defines initial specifications for the 5G network, where the novel NR technology for 5G RAN, the critical specifications for 5GC, and the technical extensions to LTE/LTE-A are introduced to fulfill 5G use cases (i.e., eMBB, URLLC, and mMTC). To primarily concentrate on novel technologies in Rel 15 for 5G RAN, this section only reviews key technological pillars for 5G NR following eMBB and URLLC functionalities, excluding enhancements from Rel 14. Each aspect provides its lessons as follows.

First, Section 2.1 introduces 5G NR, which is the main evolution from LTE/LTE-A to 5G. Without disrupting the current LTE/LTE-A systems, non-standalone and standalone architectures are accepted for the initial 5G. The non-standalone architecture combines 5G NR and existing LTE/LTE-A networks, which is a stepping stone toward a full 5G deployment. In contrast, the standalone architecture enables 5G NR systems to operate individually. The CU-DU split architecture and CU-CP/CU-UP functional split are further specified by 3GPP Rel 15 for the 5G RAN infrastructure to enhance the system operation. Although O-RAN is out of the scope of this work, it is worth mentioning that O-RAN allows more disaggregation with a CU, a DU, and a radio unit, along with two RAN intelligent controllers.

Subsequently, the eMBB functionality with a high spectrum specified in 3GPP Rel 15 is reviewed in Section 2.2. Compared to LTE/LTE-A, 5G NR allows for extremely high-frequency bands and large channel bandwidths. Table 2 summarizes the applicable frequency bands, newly defined NR bands, supported channel bandwidth, various applicable SCS, and multiple numerologies for both FR1 (sub-6 GHz) and FR2 (mmWave) classifications. Accordingly, the NR frame structure is appropriately designed to be compatible with such high-frequency characteristics. The choice of operation bands between FR1 and FR2 should consider the quality of service (QoS) requirements under the trade-offs between the geographical area coverage and high-speed DL with low latency. Notably, in Rel 15, the FRs were designated as licensed bands. To help mobile network operators deliver 5G with better performance by alleviating spectrum constraints, 3GPP Rel 16 introduced NR for unlicensed spectrum (NR-U). This extension

made advanced features of 5G NR available to unlicensed spectrum globally.

In Section 2.3, we review eMBB functionality with the massive MIMO technology specified in 3GPP Rel 15, where up to 256 antenna elements at each gNB and up to 32 antenna elements at each terminal are supported. In addition, 5G NR massive MIMO provides the capability of multilayer data transmission using multiple antenna ports and a flexible framework for TDD and FDD modes. Moreover, 5G NR massive MIMO enables a hybrid beamforming architecture by combining digital beamforming with analog beamforming using subarrays to overcome the considerable propagation loss in the high-frequency bands, especially in FR2. Beam management in 5G NR is performed by introducing new RSs and beam-functional elements, briefly summarized in Table 3. Finally, the descriptions of DL and UL transmit beamforming are also introduced. It is noteworthy that while enhancing spectral efficiency was the primary KPI for massive MIMO in Rel 15, this objective did not appear to be the main focus in Rel 16. In 3GPP Rel 16, the primary emphasis shifted towards improving the stability of MIMO connections and implementing a more resilient recovery process for beam failure.

In Section 2.4, we emphasize that 3GPP Rel 15 lays a foundation for URLLCs, where the KPIs for latency and reliability are clarified. Accordingly, a general URLLC requirement is defined with the reliability of five nines and the UP latency of 1 ms associated with a reference payload size of 32 bytes. For other specific IoT applications, latency can be relaxed to 10 ms or longer, whereas reliability can be guaranteed at six nines or even higher. Moreover, the KPI of mobility interruption time is almost free of latency for intra- and inter-frequency handovers of intra-NR mobility. While Rel-15 laid the groundwork for the URLLC use case, it did not incorporate all the necessary elements. Subsequently, Rel 16 built on the improved URLLC components, particularly focusing on latency reduction.

Notably, although 3GPP Rel 15 introduced 5G NR, only enhanced V2X was supported using LTE-A sidelink (SL) transmissions. V2X refers to the communication established between a vehicle and a variety of other vehicles, network entities, or roadside units with primarily facilitating URLLC requirements. In addition to the efforts made in Rel 14 to facilitate V2X services using LTE/LTE-A, the Rel 15 initiative provides detailed service requirements for enhanced V2X aimed at augmenting 3GPP support for V2X scenarios [60]. Therefore, 3GPP Rel 15 with enhanced V2X may not fulfill the IMT-2020 requirements for 5G services. To overcome this hurdle, 3GPP standardized NR V2X with NR SL transmissions in Rel 16.

3. Release 16

3GPP completed Rel 16, which greatly expanded the research of 5G to new services, spectra, and deployments and unlocked several new 5G opportunities beyond the traditional vertical mobile and broadband ecosystem. The most notable features mentioned are enhancements in NR non-orthogonal multiple access (NOMA) systems, integrated access and backhaul networks (IAB), positioning functions, latency reduction, energy saving for UE, NR V2X communication, and NR-U.

3.1. NR Non-Orthogonal Multiple Access

Orthogonal multiple access (OMA) systems, including time-division multiple access, frequency division multiple access (FDMA), and code division multiple access (CDMA), are commonly used in the majority of conventional cellular networks. A time-division multiple access-based system necessitates precise temporal synchronization, which is very difficult for UL transmission because the information for each user is provided in nonoverlapping time slots. In FDMA systems, such as orthogonal FDMA, the data for each user are allotted to a subset of subcarriers. In addition, a CDMA scheme uses user-specific codes to differentiate multiple users from the same channel.

However, none of these methods can fulfill the stringent requirements of future radio access systems [61]. Specifically, the number of terminal devices is restricted due to the limited radio resources. In addition, the user scheduling with the expense of control signaling overhead to guarantee the orthogonality causes high latency. Hence, NOMA has been introduced as an alternative to OMA technology [62]. Compared to OMA schemes, the fundamental mechanism of NOMA is to multiplex various data streams using the same radio resources while using a multiuser detector at the receiver to separate different signals from multiple users. Thus, constraints of orthogonal resource allocation and user scheduling time are relieved. The theoretical analysis and numerical results in [62, 63] demonstrated that NOMA outperforms OMA concerning energy efficiency, system throughput, and capacity.

The NOMA-based communications were first specified in 3GPP Rel 14 LTE to support DL and UL multiuser superposition transmission (MUST), in which NOMA can achieve better system throughput than that of OMA [64]. The NOMA-based MUST was developed for 5G NR, especially for UL transmission. Compared to OMA, the scheduling-based NOMA can achieve better spectral efficiency in several UL transmission scenarios, such as big-data communications. However, in SPC, grant-free methods must be implemented to reduce the latency, signaling overhead, and power consumption of terminal devices. Thus, NOMA-based UL grant-free communication has become a vital focus in 3GPP Rel 16 [65].

Fig. 8 compares the procedures of an UL grant-free NOMA communication based on [66] and those of a scheduling-based transmission. As depicted in Fig. 8(a), the UE must conduct a four-step random-access procedure, namely the random-access channel (RACH), with a base station (BS) before being activated for data transmission. The details of these initial random-access steps are analyzed in Subsection 3.4. Subsequently, the terminal device must send a scheduling request to the BS and wait for an UL grant before being able to transmit an information packet. The preprocessing steps between the BS and UE before the UL data transfer result in extensive control signaling overhead and a long delay in the data traffic.

In contrast, the grant-free NOMA provides autonomous and decentralized transmissions without dynamic grants. Therefore, the proposed grant-free access can resolve the signaling

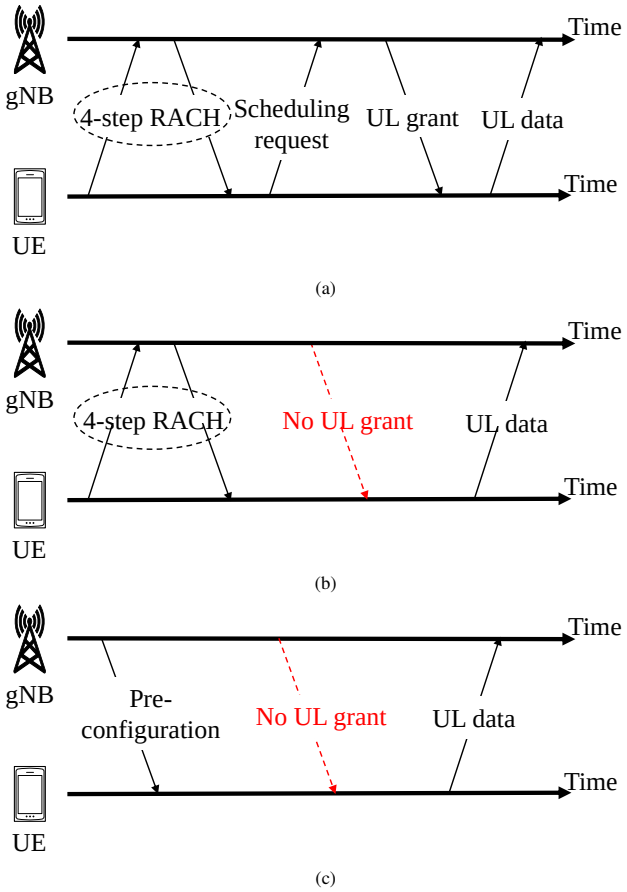


Figure 8: Procedure descriptions of (a) the grant-based, (b) RACH-based grant-free, and (c) RACH-less grant-free UL NOMA-based communication.

overhead and latency problems. According to the existence of the random-access stage, the grant-free transmission can be categorized into RACH-based and RACH-less grant-free UL communications [67], as illustrated in Fig. 8(b) and (c), respectively. Specifically, all user signals are synchronized by the random-access procedure in RACH-based grant-free transmission. This scheme can reduce the signaling overhead and benefit signal detection at the receiver. In RACH-less grant-free NOMA, both random access and grant acquisition are not performed by the users; thus, the energy consumption of devices can be reduced while receiving no preconfiguration packet. However, the RACH-less method may cause substantial complexity for the latter signal detection.

The signal preprocessing at the transmit side has become a key technical aspect in standardizing future NOMA-based communications. In MUST, transmit processing is critical because it alleviates the complexity burden of multiuser detection (MUD) and prevents the overloading phenomenon at the receiver. Several schemes have been proposed to support the UL transmission and can be categorized into bit-level and symbol-level operations [65], as depicted in Fig. 9. In Fig. 9, yellow-shaded blocks are legacy processing methods from the existing NR design, green-shaded modules are newly presented technologies for NOMA, and pink-shaded blocks are optional. The UL NOMA processing architecture design combines resource mapping and the interleaving or scrambling pattern at two levels (bit- and symbol-level operations). The bit-level operations exploit the bit redundancy in low-rate forward error-correction coding to differentiate multiple users and closely relate to channel decoding. In contrast, the symbol-level operations focus on symbol spreading and scrambling patterns. Within the UL NOMA scope, several academic studies of bit- and symbol-level schemes have been actively investigated to support grant-free transmission and are summarized in Table 6.

Recently, numerous NOMA technologies have been proposed to facilitate non-orthogonal resource allocation among the UE, and they can be typically classified into two main categories: power- and code-domain NOMA mechanisms. In power-domain NOMA (PD-NOMA) [68, 81, 82], the signals of users are multiplexed in the power domain with the same time and frequency resources. On the receiving side, multiple users are distinguished in the MUD (i.e., successive interference cancellation (SIC)) by exploiting the channel gain difference between multiplexed users. The greatest advantage of PD-NOMA is that it does not rely on the instantaneous CSI of a strongly frequency-selective fading channel; therefore, the system latency reduction can be obtained in a practical multicell NOMA system. The PD-NOMA can offer a higher outage performance and a better ergodic sum rate than OMA methods [83]. However, resource allocation and user scheduling problems are generally NP-hard in PD-NOMA transmission designs. The energy-efficiency maximization problems are the summation of multiple nonconvex subfunctions [84, 85], which yields a remarkably higher complexity than that in the OMA system. The code-domain NOMA schemes have become preferable techniques to reduce the computational complexity to overcome this obstacle.

The concept of code-domain NOMA is inspired by the conventional CDMA scheme in which multiple user-specific signature patterns are used to spread modulated symbols. Unlike CDMA, the trade-off of orthogonality in many cutting-edge NOMA technologies is to achieve higher system throughput and accommodate as many terminal devices as possible. These are the key goals of 3GPP Rel 16 [65]. The interleave-division multiple access (IDMA) has recently been widely studied in academia and industry [69, 86]. Following the basic principle of CDMA, IDMA offers a more robust solution to avoid the rate loss suffered by the intercell interference and fading effect in CDMA by exploiting several interleaves [69]. In addition, some attractive features of IDMA, including the flexible rate adaptation, power efficiency, and frequency diversity, are included in [87, Ch. 13]. Compared to CDMA, the spreading sequences are replaced by multiple low-rate coding interleaves, allowing a low-complexity elementary signal estimator algorithm, called chip-by-chip decoding strategy, to be employed at the receiver [69].

The other bit-interleaving scheme is an interleave-grid multiple access (IGMA) [70, 88]. Unlike IDMA, which only exploits bit-level interleaves, the IGMA scheme combines different grid-mapping patterns and interleaving mechanisms to improve the degree of freedom for user separation. In IGMA, the transmission bits are randomly distributed, and then the zero

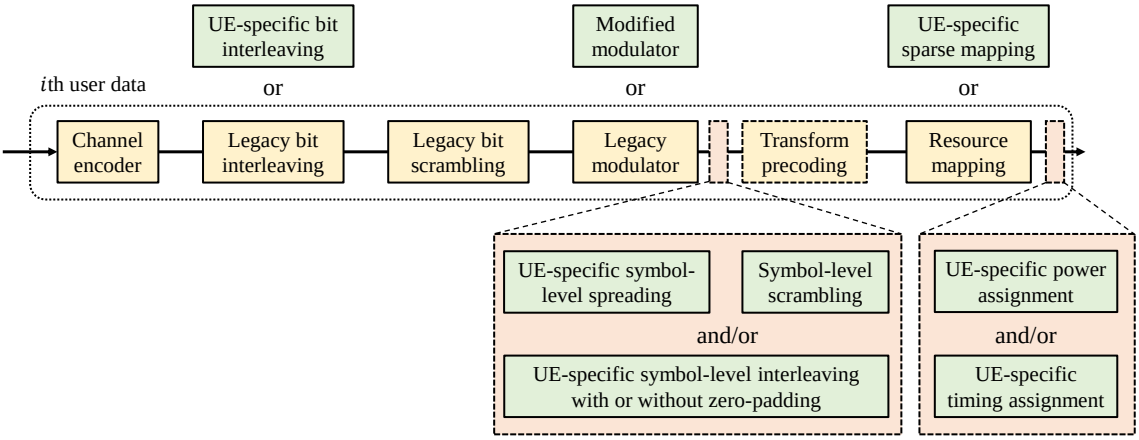


Figure 9: UL NOMA transmit processing architecture studied in 3GPP Rel 16 [65] (modules with dash blocks are optional).

padding-based mapping process and symbol-level interleaving are exploited to expand the dimensions of user multiplexing [70, 89]. The interleaving, in which the symbol sequence order is randomized, contributes to intercell interference reduction and mitigates the frequency-selective fading phenomenon. Based on the property of interleaving, the low-complexity elementary signal estimator algorithm can be applied at the receiver.

The other bit-level processing-based NOMA technique is a low code rate and signature-based shared access (LSSA) [71], which supports the UL transmission of mMTC. By exploiting a user-specific signature pattern comprising an RS, complex/binary sequence, and permutation pattern of a short-length vector, LSSA can multiplex user signals at both the bit and symbol levels. In addition, LSSA can optionally have a multicarrier variant operating at a wider bandwidth to exploit the frequency diversity and achieve latency reduction. Furthermore, low code rate spreading (LCRS) is a low coding rate-based technique that spreads the user data bits [72], where the maximum coding gain can be accomplished by implementing channel coding and bit-level repetition.

Several existing code-domain NOMA schemes are based on short-spreading sequences in symbol-level operations to improve connectivity. Regarding the sequence densities, these short sequence-based methods can be further classified into two subclasses: sparse and dense linear spreading. The main idea behind sparse-sequence-based spreading is the sparsity pattern in structured codebooks to facilitate user separation and reduce multiuser interference. The number of nonzero elements in each codeword equals the number of orthogonal resources and demonstrates the diversity order.

Sparse-coded multiple access (SCMA) [73] was introduced as the first short, sparse-spreading NOMA scheme to overcome the overloading issue in a dense network. The design aspect of SCMA was based on a predefined multidimensional codebook that creates space diversity by shuffling the signal components from multiple radio resources [90, 91]. In addition, the sparse structure of SCMA codewords can be associated with the message passing algorithm in the latter user data stream detection at the receiver [92]. However, the sparsity of the SCMA structure is not always held, especially when the number of pieces of UE increases significantly and a single carrier is performed. This outcome results in a massive delay and complexity of the latter MUD. Pattern-division multiple access has become a NOMA enhancement to meet the exponentially growing demand for large-scale mobile communications [74]. In pattern-division multiple access, irregular sparse signatures, which are the joint design of diversity and power disparities, can be exploited by a low-complexity SIC algorithm at the receiver [93, 94, 95]. Additionally, low-density signature vector extension is another improvement in the low-density spreading family [75], which can achieve a high diversity gain by merging two user-specific signature patterns into an extended vector [96].

In contrast, the short, dense sequence-based spreading schemes, such as multiuser sharing access (MUSA) [76], non-orthogonal coded multiple access (NCMA) [77], and non-orthogonal coded access (NOCA) [78], use multiple low cross-correlation sequences to ensure high MUD accuracy with a low data storage requirement. In MUSA, the data symbols are spread with multiple complex-valued sequences selected from a resource pool [97]. The spreading sequences in MUSA have a low cross-correlation to facilitate the nearly optimal SIC detector at the receiver. Moreover, NCMA is also based on low-correlation-spreading sequences, referred to as non-orthogonal cover codes, which are obtained by finding solutions to a Grassmannian line packaging problem [98]. In the NCMA scheme, parallel interference cancellation is employed to recover multiplexed signals. In addition, NOCA is another short, dense spreading scheme in which the non-orthogonal sequences are operated over the time and frequency domains [78].

In short-spreading sequence-based NOMA techniques, synchronization is required between the received signals, whereas asynchronous transmission can be supported in the long sequence-based category. Resource-spread multiple access (RSMA) with its single-carrier variant, which allows grant-free transmission, is a good candidate for asynchronous access [79].

Table 6: Summary of transmit processing technologies for NR NOMA.

OL	Scheme	Company	Mechanism		Descriptions and highlights	MUD type
Bit-level	PD-NOMA [68]	NTT DoCoMo	Power-based		The signals are multiplexed in the power domain. This scheme design can be decoupled into two subproblems: power allocation and user scheduling. These problems are nonconvex and require considerable computational complexity.	SIC
	IDMA [69]	Nokia	Interleaving	Low coding rate	IDMA can mitigate the intercell interference and fading effect of CDMA. It can exhibit attractive properties, such as flexible rate adaptation, power efficiency, and frequency diversity.	ESE
	IGMA [70]	Samsung	Interleaving		A grid-mapping pattern and interleaving mechanism are employed to enhance the user separation performance. The symbol-level interleaving mitigates intercell interference and frequency selectivity fading.	ESE
	LSSA [71]	Samsung	Scrambling		LSSA was proposed to support UL asynchronous large-scale transmissions by multiplexing the user-specific signature patterns. It can be used to achieve frequency diversity and lower latency.	SIC
	LCRS [72]	Intel	Spreading		The maximum coding gain can be accomplished by implementing channel coding and bit-level repetition.	SIC
Symbol-level	SCMA [73]	Huawei	Spreading	Short and sparse-spreading sequence	Coded bits are mapped to the modulation symbols based on a sparse multidimensional codebook. When a single-carrier transmission is performed, numerous users may result in substantial processing complexity and delays at the receiving MUD.	MPA
	PDMA [74]	CATT	Spreading		Irregular sparse signatures are employed to adapt to a massive number of users. Multiple domains, including space, code, power, and combinations of these, are expanded to enhance the multiuser separation performance.	MPA, SIC
	LDS-SVE [75]	Fujitsu	Spreading		Unlike SCMA and PDMA, which employ separated user-specific signature patterns, LDS-SVE combines two signature vectors into a more significant spreading vector. Thus, this scheme can exploit a higher diversity gain.	MPA, SIC
	MUSA [76]	ZTE	Spreading	Short and dense spreading sequence	Several short complex-valued sequences are employed. One user can transmit various symbols in multiple short sequences selected from a spreading sequence pool. The existence of an imaginary part helps expand the resource pool while achieving a low signal cross-correlation.	SIC
	NCMA [77]	LG Electronics	Spreading		The interference between two users can be predicted based on non-orthogonal cover codes, which are obtained by solving the Grassmannian line packaging problem. Several multistage spreading sequences based on non-orthogonal cover codes can be applied to improve the system throughput.	PIC
	NOCA [78]	Nokia	Spreading		The spreading mechanism of NOCA is operated in both the frequency and time domains. The spreading sequences in NOCA are designed based on multiple spreading factors that support adapting to user detection accuracy requirements.	SIC
	RSMA [79]	Qualcomm	Scrambling	Long spreading, scrambling sequence	Codewords for each user can be spread in all available time and frequency domains. The RSMA can achieve high system reliability by employing low-rate forward error-correction codes and scrambling sequences with low correlation. Two kinds of schemes, single- and multicarrier RSMA, are implemented based on different scenarios.	SIC
	RDMA [80]	MediaTek	Interleaving		In RDMA, modulated symbols of each user are repeatedly transmitted, and various cyclic-shift numbers are indexed for multiple iterations. The RDMA reduces the signaling overhead because the user-specific interleaver and scrambler are not necessarily required.	SIC
	GOCA [80]	MediaTek	Spreading		The GOCA is an improvement of the RDMA. It employs multiple group orthogonal sequences with a two-step structure to spread the modulated symbols over shared time and frequency resources.	SIC

Similar to scrambling-based LSSA, in RSMA, the user codewords are spread over the available time and frequency domains; thus, RSMA can achieve full diversity [79]. Moreover, RSMA can achieve high system reliability by employing low-rate forward error-correction codes and scrambling sequences with low correlation [82]. In [99], the multicarrier RSMA is also emphasized as an efficient solution for low-latency access.

In the long-spreading sequence-based family, repetition-division multiple access (RDMA) has been proposed to exploit both time and frequency diversity with the cyclic-shift repetition of multiple modulated symbols [80]. RDMA deploys symbol-level interleaving based on cyclic-shift repetitions, whereas IDMA focuses on bit-level interleaving. RDMA also reduces signaling overhead compared to IDMA and RSMA

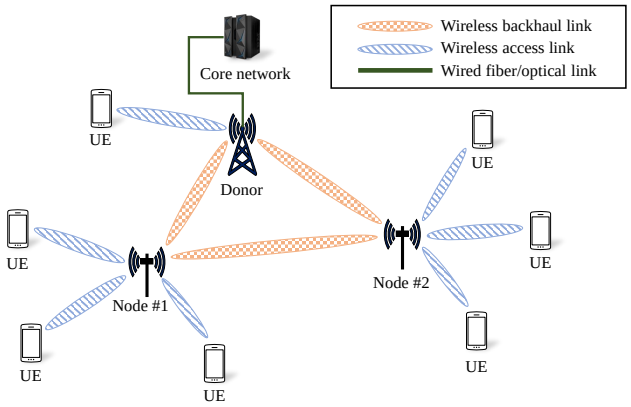


Figure 10: Wireless backhauling network architecture.

because the user-specific interleaver and scrambler are not required. Additionally, group orthogonal coded access is an enhancement of RDMA [80] and is a two-stage technique that generates multiple grouped orthogonal sequences. These sequences are spread over shared time and frequency resources.

Numerous transmission schemes were conducted by different companies, as shown in Table 9, each developing their own NOMA schemes. Those schemes have different advantages and disadvantages and can be categorized into the following groups: scrambling-, spreading-, and interleaving-based NOMA schemes. Scrambling-based NOMA offers low complexity, making it suitable for conveying control channels. However, its performance is degraded when user channels differ significantly. Spreading-based NOMA exhibits stronger interference suppression due to inherent user separation provided by spreading codes. This robustness to diverse channel conditions facilitates its application in data channels and scenarios with varied user complexities. However, the increased complexity associated with spreading operations necessitates trade-offs compared to scrambling-based approaches. Finally, interleaving-based NOMA leverages channel diversity for user separation, demonstrating robustness to frequency-selective fading and enhancing multi-user diversity. This makes it an attractive choice for URLLC applications [88].

3.2. Integrated Access and Backhaul Networks

In a service network, the backhaul links a remote area (i.e., RAN) and a central management node (i.e., the core network). Due to the requirement for high-speed Tbps rates and reliable information transportation, fiber backhauling has typically been the leading technology for this application. However, wireline deployment requires a remarkable investment in hardware installation, and fiber communication is not an option for future technologies. With the projected densification of networks and introduction of new technologies, such as unmanned aerial vehicles (UAVs) [100, 101], massive MIMO, high altitude platform (HAP) stations, and beamforming, wireless backhauling has become a mandatory, integral part of 5G communications due to its significant reduction in cost and infrastructure.

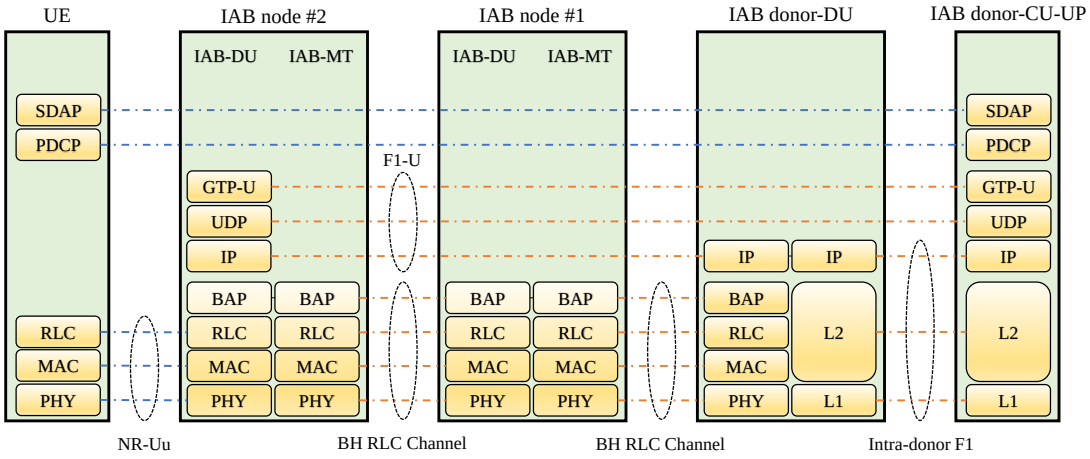
Fig. 10 demonstrates an architecture combining wireless backhauling links with fiber installation. A gNB-controlling donor can connect to mobile users via access links and communicate with other gNB distributing nodes via backhaul chains, such as a small BS or a UAV. These nodes provide communication access to the local terminal devices and connect to other small BSs in different cells. These nodes form a wireless relay network that extends the coverage of the donor station connected optically to the 5GC network.

The wireless backhaul study was first presented in 3GPP Rel 10, known as LTE relaying [102]. The LTE spectrum that cannot be utilized for backhauling purposes is expensive; thus, the existing LTE relay deployments cannot meet the commercial criteria. The 5G licenses have been in commercial use over the past few years to facilitate a larger spectrum, such as mmWave, which provides a broader bandwidth for wireless backhauling. With a higher-frequency bandwidth, the operators can partition the radio resources into wireless access and backhaul. In other words, the available resources must be efficiently allocated to balance the user network access and the backhaul traffic transmission of the BS to the core network.

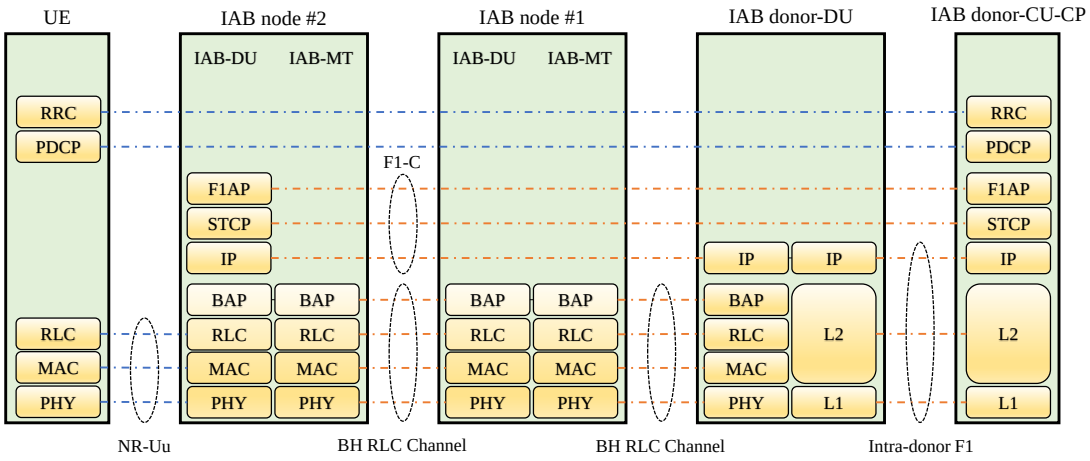
Initially, the fixed access backhaul network was proposed as an allocation solution in which the resources for various access and backhaul links are equal across all BSs [103]. We let T_A and T_B be the allotted time for the access and backhaul links, respectively, and assume that the total available time is 1 s without loss of generality. Then, the constraints of resource allocation in a fixed access backhaul network are $T_A \leq \delta$ and $T_B \leq 1 - \delta$ for $\delta \in [0, 1]$. However, the allotted time for access and backhaul links for each BS maintains the global partition; thus, this results in limitations on the end-user QoS, including the transmission rate and latency.

The IAB has been standardized in 3GPP Rel 16 [104] and recognized as a cost-effective replacement for resource management problems as it allows a more flexible partition of access and backhaul resources that enhance the QoS of the entire communication network. Although the total available resources for access and backhaul links of each BS are still fixed, no global partition between two links is maintained in an IAB network. In other words, the allotted time constraint is replaced as $T_A + T_B = 1$; therefore, IAB helps overcome an overloaded situation. Specifically, when the BS has significantly more connections to the terminals than usual, it is more beneficial to employ IAB, which uses the available backhaul resources to access links. Furthermore, IAB can significantly reduce dependence on the wired backhaul deployment compared to that of fixed access backhaul [103]; thus, IAB is more advantageous to meet the dynamic traffic demand of mmWave communications.

Fig. 11 describes the IAB user and control protocol stack according to [104]. The overall IAB architecture is based on the functional split [105], where the IAB donor is split into a CU and multiple DUs. A DU terminates the lower layer protocols, including RLC, MAC, and the PHY protocol, while a CU, apart from PDCP, provides support for the higher layers of the protocol stack, such as service data adaptation and RRC protocol in UP and CP, respectively. The IAB donor directly connects to the core network, employs the DUs to help serve other IAB nodes and UEs, and extends the data transmission coverage. Two *F1* interface types, *F1-U* and *F1-C*, convey data in the UP



(a)



(b)

Figure 11: Descriptions of a protocol stack for (a) IAB UP and (b) IAB CP according to 3GPP Rel 16 [104].

and CP, respectively.

Regarding the functional CU–DU split architecture, the general packet radio system tunneling protocol and user datagram protocol packets are conveyed between the DU part of an IAB node and the UP of the CU over an *F1-U* interface. Similarly, for the CP, the corresponding stream control transmission protocol and *F1* application protocol packets are transferred via the *F1-C* interface. In addition, an IAB node may have a backhaul link to the IAB donor through several connections to other IAB nodes. Therefore, a multihop backhauling system can be formed in the IAB network, as depicted in Fig. 11. Each IAB node has a DU and a mobile termination (MT) functionality. The MT is used to communicate with a parent DU of a parent IAB node or an IAB donor. The DU functionality is to connect to an MT of a child IAB node.

The CU–DU split architecture is motivated by the time-critical functionality division between two units. For example, scheduling, quick retransmission, and segmentation operate from the DU side, which is close to the antenna station. Other less time-dependent tasks are realized in the CU part. In addition, the external interface (i.e., Xn), which conducts connections between a set of gNB-CU nodes, can be terminated in the CU part; thus, the extra complexity is reduced on the DU sides. Furthermore, the split structure also enhances the security protection of end-to-end communication between the CU and a separate UE by supporting the centralized termination of PDCP, as displayed in Fig. 11. Besides the CU–DU split standard, a group of architectures presented in [106] has no split and offers a nested tunneling connection between an IAB donor and IAB nodes.

3.3. NR Positioning

Terminal wireless positioning has become a significant application in several areas, such as location-based services, healthcare sensor management, and UAV operations. The positioning function has been supported in the cellular network since Rel 9, and radio access technology-independent techniques (i.e., global navigation satellite system (GNSS), blacktooth,

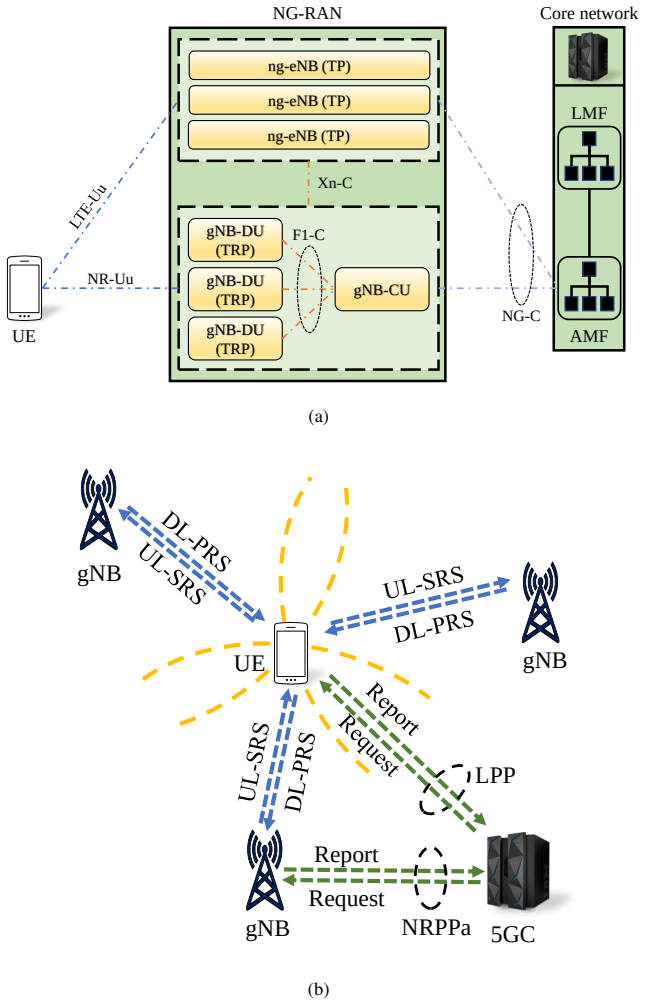


Figure 12: Description of (a) a typical UE positioning architecture, and (b) TDoA procedure for NR Rel 16.

barometric pressure, and Wi-Fi signal strength) have been supported on LTE carriers since Rel 15. Although GNSS is a successful positioning innovation, it only applies to numerous outdoor scenarios. The 5G technology improves outdoor positioning accuracy and provides exact indoor positioning performance [107]. The positioning techniques in 5G are predominantly based on 4G architecture, with some innovative modifications adapting to the existence of the 5GC.

In addition, 3GPP Rel 16 proposed the UE positioning architecture, as demonstrated in Fig. 12, that applies to the NG RAN [108]. As illustrated in Fig. 12(a), the gNB and enhanced LTE eNB are the NG RAN nodes receiving the positioning requests from a location management function (LMF), convey request signaling to the target UE, perform UL positioning measurements, and handle information transfer between the NR RAN and access and mobility function in the core network. As depicted in Fig. 12(b), two main protocols are employed. The LTE positioning protocol defines the procedures to exchange signals between the UE and several gNB and enhanced LTE eNB transmission points in 4G. It has been extended to support signaling between the UE, gNB transmission reception points (TRPs), and LMF in 5G positioning [108]. Additionally, the NR positioning protocol annex [109] covers the positioning measurement control signaling between NG RAN BSs and the LMF.

In positioning architecture, an LMF determines a positioning technique based on the UE location and QoS requirements for latency and accuracy. In addition, Rel 16 introduces some RAT-dependent techniques, including the DL/UL time difference of arrival (TDoA), DL angle of departure (AoD), UL angle of arrival (AoA), multicell round trip time (RTT), enhanced cell identity to meet the more stringent performance requirements for commercial use cases. For example, the targets of positioning error are below 3 m indoors and 10 m outdoors at 80% reliability for commercial use cases while maintaining less than 1 s for the end-to-end latency [110]. High-accuracy applications, e.g., V2X, necessitate even stricter goals of less than 1 m error for network-based 3D positioning, necessitating high reliability and support for challenging environments [110]. Table 7 summarizes critical performance requirements for 5G NR.

In practice, timing-based methods are the most commonly used among the mentioned positioning techniques. Especially for electronic warfare, the TDoA schemes offer more accurate measurements than conventional triangulation methods [111]. The prior time-based localizing techniques, the observed TDoA and UL TDoA, were defined in the 3GPP LTE Rel 9 and Rel 11, respectively. These geolocation methods are standardized in NR Rel 16 and are called the DL TDoA and UL TDoA. The TDoA localization uses hyperbolic trilateration (or multilateration), which is visualized by three dashed yellow hyperbolic curves in Fig. 12(b), to perform the terminal position approximation. The multilateration in DL TDoA can be performed based on multiple measurements of the time of arrival of the RSs from different gNBs to the terminal.

The mechanism of the UL TDoA positioning technique is similar to that of DL TDoA, except that it uses the UL RS transmitted from the UE to the neighboring gNBs. The UL TDoA calculation can be performed at the location measurement units, which are any standalone units placed either inside or outside the gNBs [112]. The information received at these location measurement units is then conveyed to the core network via the NR positioning protocol annex to compute the TDoA. Compared with the DL TDoA, the localizing measurements are not executed at any UE; thus, the computational complexity is significantly reduced for each terminal.

Because the accuracy of time-based measurements rises with the signal bandwidth, using wideband or ultra-wideband signals for ranging applications is appealing. As a result of its capacity to resolve closely spaced multipath components, ultra-wideband signaling has the potential to achieve high-precision localization even in congested situations. These time-based schemes typically operate based on estimating the arrival time of the direct signal. However, in a contaminated environment affected by multipath propagation, noise, and interference, the first incoming signal is possibly a multipath component and hinders the TDoA or time of arrival estimation schemes [113]. When the transmission time is unknown beforehand, the local BSs can convey messages containing processing time to estimate the RTT instead [114].

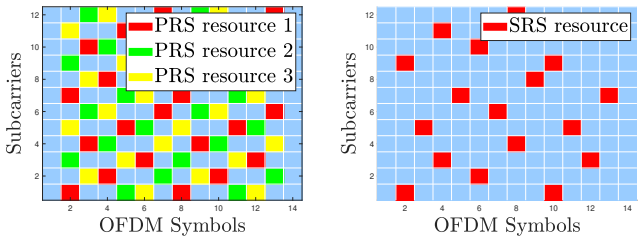
Table 7: NR positioning requirements in Rel 16 and beyond [110].

Performance Metric	Regulatory Minimum	Commercial Indoor	Commercial Outdoor	High Accuracy (e.g., V2X)
Horizontal accuracy	≤ 50m (80% UEs)	≤ 3m (80% UEs)	≤ 10m (80% UEs)	< 1m (> 95% service area)
Vertical accuracy	≤ 5m (80% UEs)	≤ 3m (80% UEs)	≤ 3m (80% UEs)	3D positioning: 10m to < 1m (80% scenarios)
Latency	< 30s end-to-end	< 1s end-to-end	< 1s end-to-end	Low latency critical
Additional requirements	-	-	-	High reliability, 200km/h mobility

The RTT has been specified in 4G LTE systems but is only restricted to serving cell or indoor scenarios. In 5G NR Rel 16, the multicell RTT method has been standardized as the measurement possible from neighbor BSs in different cells. The mechanism of multicell RTT is that the UE sends a requesting UL RS and receives multiple responding DL RSs from the nearby BSs. The RTT estimated from those access communications is used to compute the distance between the UE and neighbor BSs. Similar to the ideas of TDoA techniques, the position of UE can be inferred by applying a hyperbolic trilateration procedure. Both DL and UP RSs are employed for multicell RTT so that the resource overhead is higher than that of the TDoA. However, the time synchronization errors of TDoA can be resolved [115, 116].

Although the multicell RTT method is advantageous in time synchronization, the angle-based solutions, including DL AoD and UL AoA, are robust high-resolution positioning methods in 5G mmWave applications and have been studied for decades. For DL AoD, the RS received power (RSRP) is generated based on the DL RSs transmitted from multiple transition points with beam sweeping to the UE. The RSRP report is transferred from the UE to the local core network via LTE positioning protocol packets. The AoD based on RSRP can be estimated by applying the fingerprinting algorithm [117]. Compared to the DL AoD, the UL AoA is the most popular angle-based positioning method and is more efficient because the angles of the incoming signals can be estimated directly at the gNBs.

Various techniques have been proposed for array signal processing-based UL AoA estimation and are classified as subspace-based methods [118, 119], sparsity-minimizing methods [120, 121], and the maximum likelihood method [122]. Subspace-based techniques, such as multiple signal classification [118] and signal parameter estimation via the rotational invariance technique [119], can achieve a higher estimate resolution with a lower computational burden. Furthermore, the direction estimation algorithms can be applied as stand-alone and hybrid solutions with other RAT-dependent techniques, such as the time-based estimation (i.e., TDoA); thus, positioning performance is significantly enhanced in 5G applications. However, the most challenging positioning problem is the existence of non-line-of-sight signals. With the assumption of far-field sources, the multipath signals hinder the implementation of the subspace-fitting techniques due to the rank deficiency of the signal covariance matrix [123, 124]. Several dimensionality reduction schemes, such as spatial smoothing [125] and algorithms similar to the signal parameter estimation via the rotational invariance technique [126, 127, 123, 128, 129] have been suggested to overcome this obstacle.



(a) (b)

Figure 13: Examples of symbol allocation of the (a) DL positioning RS and (b) UL sounding RS in a physical resource block.

Although multiple existing RSs, such as synchronization signals or CSI-RS, can be opportunistically exploited for localization estimation, they have several drawbacks. Initially, when the signals from many BSs collide in both the time and frequency domains, the interference from the adjacent cells prevents these RSs from identifying the number of neighbor gNBs. In addition, these RSs exhibit weak correlations due to their low resource element density and potential for their resource element pattern to not extend throughout all of the frequency-domain subcarriers. Fig. 13 visualizes two new RSs proposed in Rel 16 for DL and UP positioning. The RSs referred to as the DL PRS and UL SRS are employed in DL and UP positioning, respectively. With a high resource element density and better correlation properties [39], these RSs have been leveraged to overcome the above limitations.

Fig. 13(a) illustrates a DL PRS configuration in a comb-6 pattern with three TRPs only in one physical resource block. By muting the appropriate PRS transmission instances, multiple gNBs transmit the PRS simultaneously without interference from nearby cells; thus, hearability is achieved with the PRS muting property. Similarly, the comb-based structure has been applied for the UL SRS since Rel 15. Fig. 13(b) visualizes an example of the comb-4 UL SRS in one physical resource block. From 3GPP Rel 16, its configuration is expanded to the comb-8 pattern in one resource [39].

During the evolution of positioning technologies documented in the technical reports of Rel 16, numerous field experiments have been executed to validate their compliance with 5G NR positioning requirements. The investigation by Papp *et al.* [130] delves into TDoA-based indoor positioning within 5G small cell networks. Through the utilization of realistic simulations and advanced algorithms, the study successfully mitigates non-line-of-sight propagation, achieving remarkable positioning accuracy consistently under 3m. Meanwhile, Menta *et al.* [131] leverage an extended Kalman filter in a 5G outdoor net-

Table 8: Comparison of 5G positioning approaches in existing studies.

Ref.	Input Data	Environment	Error Range
[130]	TDoA measurements	Indoor	3 m (can be improved)
[131]	AoA estimates	Outdoor	Sub-meter (95% probability)
[132]	AoA & TDoA	Outdoor (vehicle-to-vehicle)	Lateral: < 1 m, longitudinal: < 0.5 m (distance constraints)
[133]	RSRP & AoD	Urban	2.1 m (line-of-sight), 8.4 m (non-line-of-sight)
[134]	Various beam data	Indoor and Outdoor	Sub-meter (median squared error)

work, showcasing their prowess by attaining sub-meter 2D positioning accuracy. Furthermore, the study by Kakkavas *et al.* [132] navigates the realm of vehicle-to-vehicle positioning, ensuring alignment with 5G NR requirements for platooning and overtaking scenarios. Malmström *et al.* [133] utilize RSRP and AoD information to estimate UE positions with an impressive 80% accuracy within 10 m. Additionally, the authors in [134] introduce a data-driven approach tailored for mmWave networks, achieving sub-meter accuracy while effectively addressing collinear regions. Table 8 demonstrates diverse techniques for 5G positioning, targeting both indoor and outdoor environments.

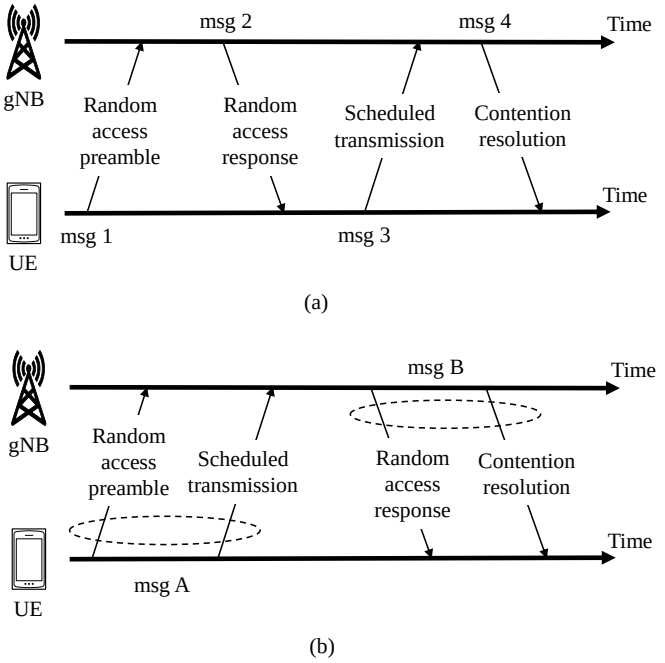


Figure 14: Evolution of the random-access procedure from (a) a four-step RACH in Rel 15 to (b) a two-step RACH in Rel 16 [135].

3.4. Latency Reduction

A typical design of a random-access procedure on PRACH is required to fulfill the low latency requirements for several 5G applications, such as URLLC, eMBB, mMTC, and the adaptation to both licensed and unlicensed spectra. For example, the UL data must be transmitted quickly from the UE to BS in an mMTC, which covers the connectivity of a million machine-type communication devices per square kilometer. From that perspective, the two-step RACH, based on the four-step RACH proposed in 3GPP Rel 15, has been standardized in Rel 16 to accelerate the initial connection establishment between the UE and BS [135].

Fig. 14 illustrates two versions of random-access procedures in Rel 15 and Rel 16. As depicted in Fig. 14(a), the four-step RACH manifests two round-trip UE–BS cycles accomplished by elaborating on the control signaling. Initially, the UE sends Message 1 to the BS, containing a randomly selected preamble associated with the PRACH occasion. A random-access radio network temporary identity is defined by the time slot of the preamble transmission that the BS can address the UE message in the subsequent step [136]. However, other UEs can also have the same temporary random-access radio network identity; thus, both the BS and UE have no evidence of this contention at the initial step. Subsequently, the BS analyzes the temporary identities of the random-access radio network to detect the received preambles. Then, it transmits a random-access response, which carries the confirmation identifier of successfully received preambles, cell radio network temporary identifiers, timing advance, and PUSCH resource grants for transmitting Message 3 from each piece of UE [136]. Next, each piece of UE applies for timing advance and transmits Message 3 to the BS based on the assigned PUSCH resource. Message 3 includes UE identity for contention resolution. The final message of the round-trip cycle contains the received UE identity, which resolves the possible contention arising due to the same preamble transmission of multiple pieces of UE in the first step.

The mentioned procedures can take several milliseconds, resulting in significant signaling overhead on the network, increased power consumption for terminal-side signal transmission and detection, and significant latency for data transmission. The two-step RACH enhancement presented in [135, 137] can reduce latency and control signaling overhead by reducing the number of messages between the UE and BS. This reduction is achieved by combining the random-access preamble in Message 1 and the scheduled PUSCH transmission in Message 3 into a single message referred to as Message A from the UE. Subsequently, the BS sends UE Message B, a combination of a random-access response in Message 2 and the contention resolution in Message 4. Thus, only one round-trip cycle between the UE and BS is needed to finish the access procedure instead of the two cycles required in the four-step RACH, as displayed in Fig. 14(b). Additionally, the reduction of transmitted messages decreases in listen-before-talk (LBT) events in the NR unlicensed spectrum.

3.5. User Equipment Power Savings

The efficient efficiency of wireless communications relies on a balance between optimizing data transmission efficiency during network access and minimizing power consumption during idle periods. This duality translates into two distinct terminal

Table 9: Summary of power consumption.

Power state		Relative power (power unit)
PSM	Deep sleep	1 (ref.)
	Light sleep	20
	Microsleep	45
NAM	PDCCH-only	100
	PDCCH+PDSCH	300

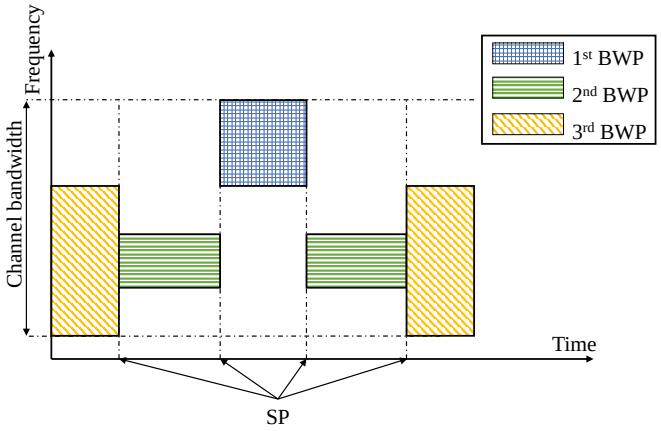


Figure 15: Bandwidth adaptation with three BWPs.

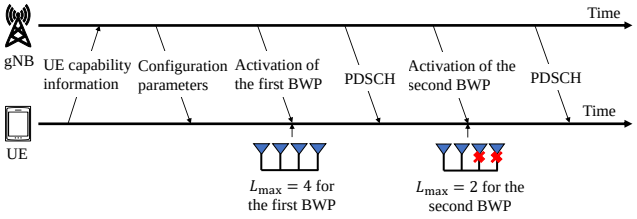


Figure 16: Example of the MIMO layer adaptation procedure.

modes: network access mode (NAM), where the terminal actively exchanges data, and power saving mode (PSM), characterized by deep sleep and minimal power use through deactivated radio and baseband components. Power-saving techniques aim to optimize this balance by maximizing NAM efficiency through bandwidth adaptations, antenna usage, and processing time while simultaneously prolonging PSM periods by minimizing the latency of mode transitions through dynamic signaling. In NR Rel 16 [138], various sleep states offer a trade-off between power consumption and wake-up latency. Deep sleep minimizes power consumption with limited baseband activity but longer wake-up times. Light sleep maintains stricter timing and faster wake-up at the cost of slightly higher power usage. Microsleep allows rapid wake-up but precludes data transmission. On the other hand, the active state encompasses PDCCH-only and PDCCH+CPDCCH configurations for data communication. The relative power per slot for each state is summarized in Table 9. We note that the power for each state is relative to that of deep sleep, which consumes a single power unit. Most UE energy is employed in the active state, and even if no transmission is scheduled, the UE still monitors the control channel, leading to power waste. Therefore, minimizing the wake-up duration is the most significant power-saving mission.

Some essential techniques for an energy-efficient 5G NR framework have been proposed. The two most typical techniques are the discontinuous reception (DRX) [139] and the bandwidth part (BWP) adaptation [140]. The DRX is a power-saving method regulating the UE to continuously receive signals with a predetermined active and sleep period under an RRC connected state. The UE with DRX monitors PDCCH periodically during the active state controlled by the DRX inactivity timer. When data are not scheduled, this parameter expires, and the UE can enter the power-saving mode in which the DRX on duration timer is employed to determine whether a control channel exists.

However, the 5G NR supports up to 400 MHz of bandwidth; thus, the terminal consumes significant power for a wide filter and a high-speed analog-to-digital (ADC) module to receive wideband signals. The BWP, which is a set of contiguous resource blocks configured inside a channel bandwidth, was first introduced in Rel 15 to support UE bandwidth adaptation that reduces device power consumption. In other words, if the BWP is smaller than the system bandwidth, the UE can tune the RF bandwidth to the BWP to save energy. As illustrated in Fig. 15, the UE configured with three BWPs can switch between different BWPs at switching points. The switching from the first to the second BWP may occur for several reasons, such as the expiration of the inactivity timer [140] of the first BWP or receiving DL control information (DCI) indicating the second BWP is active.

Based on the mentioned inventions, a standardized framework has been built since the first release of 5G NR, and some improvement techniques have been introduced in Rel 16 to meet the 5G key requirement of energy-efficiency enhancements by a factor of 100 [141]. First, adaptive MIMO layer reduction is introduced to reduce the number of DL MIMO layers in some low or latency-tolerant traffic load scenarios. To this end, the maximum number of MIMO layers $L_{\max}$ for DL data reception in one serving cell is employed. Fig. 16 illustrates a procedure of MIMO layer adaptation with two BWPs. The UE may inform the BS of its capability information, indicating the $L_{\max}$ values regarding two BWPs via an RRC message. The UE receives the second RRC message mentioning configuration parameters, comprising the $L_{\max}$ values assigned for each BWP. As depicted in Fig. 16, the UE activates up to $L_{\max} = 4$ antennae if the first BWP is in use, whereas this number is reduced to only $L_{\max} = 2$ if the second BWP is activated. Thus, the computational resource can be dynamically reduced by skipping the channel estimation for the unused antennae and turning off the RF components.

The disadvantage of the DRX mechanism is that the UE must wake up periodically to monitor the PDCCH, even during sporadic traffic or when no data arrives. A wake-up signal (WUS) was proposed to monitor the channel control to avoid this unnecessary power consumption. The WUS is provided by a new

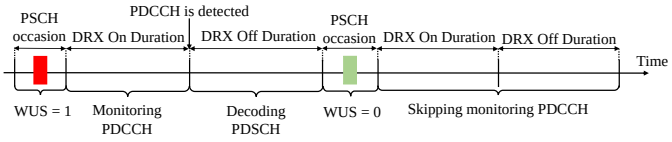


Figure 17: Wake-up signal-based power saving mechanism.

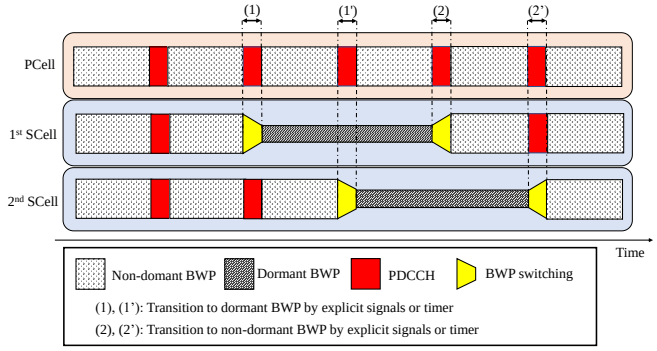


Figure 18: Example of the SCell dormant state transition based on the PCell signaling.

DCI format called DCI format 2_6 [142], which is employed to notify the power-saving information located in a power-saving channel before a DRX on duration timer for a DRX cycle, as presented in Fig. 17. Based on the binary level indicated by the WUS, the terminal wakes up to start the DRX on the duration timer and monitors the PDCCH if the WUS is set to a high level. In contrast, the WUS may become a go-to-sleep indication when set to a low level. Accordingly, the UE does not need to start the timer and can stop waiting for any PDCCH. Thus, this method efficiently regulates the active duration of a DRX cycle to reduce power consumption on top of the DRX mechanism.

Several serving cells, including a primary cell (PCell) that performs network access managing procedures and multiple secondary cells (SCells), are aggregated simultaneously to serve the UE through the CA operation to support a larger bandwidth. The CA has supported LTE systems to activate or deactivate SCells rapidly, depending on the traffic conditions. However, the deactivating and reactivating state transition of SCells consumes considerable power due to some crucial procedures, such as beam management or CSI measurement. In Rel 15, a new SCell state, referred to as the dormant state, is introduced so the terminal can stop decoding the PDCCH in that SCell without affecting other procedures; therefore, unnecessary power consumption can be avoided. The dormant state transition is accomplished by receiving the activation/deactivation and hibernation MAC elements [143].

In Rel 16, the SCell dormancy is reported to be implemented based on the BWP framework. Specifically, the SCell can be switched from dormant to active when data transmission occurs on that SCell. As depicted in Fig. 18, the SCell dormancy is conveyed by receiving BWP switching DCI with format 2_6 [144] on the PCell or the primary, secondary cell when the terminal is outside the DRX active time. Through DCI signaling, the UE may reduce the state transition latency so that the assigned SCells can adapt to the traffic load condition more quickly and frequently than MAC-control elements, which occupy a large portion of processing time [145]. In addition to the mentioned WUS indicators, the SCell dormancy indicators are integrated into the same DL control signaling for multiple pieces of UE to reduce the resource overhead, as illustrated in Fig. 19.

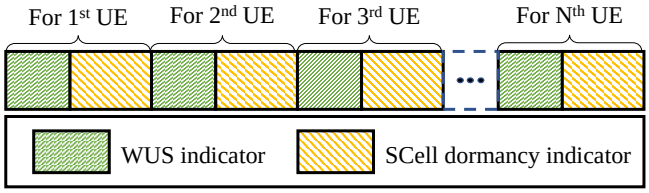


Figure 19: Structure of a DCI format 2_6.

The concept of cross-slot scheduling is introduced as a power-efficient technique for 5G NR to help the terminal reduce energy by expanding the low-power state without affecting performance. This technique informs the terminal in advance when the DCI is transmitted via the data channel (i.e., PDSCH or PUSCH appears in a slot after the slot containing the control channel, the PDCCH). When no data are scheduled, the device omits some unnecessary RF operations to buffer the PDSCH and go into a microsleep state rather than performing nonessential decoding tasks. According to [146], the time delay values between the PDCCH slot and scheduled PDSCH and PUSCH slots are denoted as K_0 and K_2, respectively. The slot number K_S allocated for the data channel is calculated as $K_S = \left\lfloor n\frac{2^{\mu_{\text{PDCCH}}}}{2^{\bar{\mu}}} \right\rfloor + K_\chi$, where n denotes the slot with the scheduling DCI, $K_\chi = \{K_0, K_2\}$ derives the slot offset from the slot where the DCI (PDSCH or PUSCH) is received, and $\bar{\mu}$ and μ_{PDCCH} represent the SCS configurations of the data channel and PDCCH, respectively. The same-slot scheduling procedure is employed when the SCS factors are identical and K_χ is zero. Typically, the same-slot scheduling is only applied when the terminal cannot determine what scheduling technique to decode the DCI. This method requires a relatively higher processing duration than cross-slot scheduling because UE must keep the RF chain on to buffer any immediate DL transmission while decoding the PDCCH.

An enhancement of cross-slot scheduling, employing the minimum scheduling offset restriction values $K_{0\,\min}$ or $K_{2\,\min}$, is also introduced to guarantee that the UE has the knowledge of the minimum slot offset before decoding the next PDCCH [42]. Specifically, when the DCI format 0_1 or 1_1 with the minimum applicable scheduling offset indicator field representing a change to the applied $K_{0\,\min}$ or $K_{2\,\min}$ is contained within the first three symbols of the nth slot, the value of delay $\overline{K_\chi}$ is given as $\overline{K_\chi} = \max\left\{K'_{\chi\,\min}\left\lceil\frac{2^{\mu_{\text{PDCCH}}}}{2^{\bar{\mu}}}\right\rceil, Z_\mu\right\}$, where Z_μ is determined by the SCS of the active BWP in the scheduling cell, and $K'_{\chi\,\min}$ denotes the currently applied minimum scheduling offset $K_{0\,\min}$ or $K_{2\,\min}$ of the active DL or UL BWP in the scheduled offset, respectively.

Table 10: All power saving techniques discussed in Rel 16

Technique	Advantages	Disadvantages
BWP management	Reduced power consumption on unused resources	Increased signaling overhead with frequent switching, potential connection delays with BWP deactivation
DRX enhancements	Extended UE sleep periods for significant power savings	Higher latency due to longer sleep periods and inactive state, reduced unnecessary wake-ups with WUS, missed network messages or delays with inactive state
MIMO layer reduction	Significant power savings on MIMO processing	Reduced data rates, especially in high-throughput scenarios
WUS	Reduced unnecessary wake-ups	Increased network complexity due to additional infrastructure and configuration requirements
SCell dormancy	Reduced power consumption on unused small cells	Potential delays in re-attaching to small cells, reduced scanning efficiency
Cross-slot scheduling	Improved overall system efficiency and UE power savings	Slight data transmission delays compared to continuous scheduling

As shown in Table 10, those aforementioned methodologies exhibit noteworthy significance; however, each approach is not exempt from inherent limitations. BWP switching may result in elevated signaling overhead during frequent adjustments, whereas the deactivation of BWP introduces the potential for re-activation delays. DRX enhancements, despite their commendable power-saving attributes, are accompanied by potential latency penalties, particularly in prolonged sleep periods and deeper inactive states. Adaptive MIMO layer reduction, while enhancing power efficiency, may compromise data rates, especially in high-throughput scenarios. While WUS proves efficient in minimizing unnecessary wake-ups, it introduces network complexity. SCell dormancy balances power savings with potential re-attachment delays and diminished scanning efficiency. Cross-slot scheduling offers power savings through UE sleep periods but may entail fewer data transmission delays than continuous scheduling. Effectively navigating these trade-offs necessitates meticulous consideration of network conditions, application requirements, and user preferences, facilitating optimal power optimization without compromising performance or user experience.

3.6. NR Vehicle-to-Everything Communications

3GPP reports [60, 147] provide a comprehensive overview of NR V2X use cases and associated requirements, dividing them into four key groups: vehicle platooning, advanced driving, extended sensors, and remote driving. Vehicle platooning involves dynamic platoon management with essential data exchange influenced by QoS. Advanced driving focuses on semi- or fully automated driving, emphasizing data sharing for safety and traffic efficiency. Extended sensors highlight sensor data exchange among vehicles, roadside units, pedestrian devices, and V2X servers to enhance environmental perception. The remote driving group enables teleoperated control for scenarios such as hazardous environments. This structured approach to automation and diverse use cases showcases NR V2X's potential to revolutionize transportation, enhancing safety, traffic efficiency, and environmental awareness through coordinated information sharing and action. In Rel 16, 3GPP sets stringent requirements for V2X communications, driving towards advanced, intelligent transportation systems, such as low latency ($<$ 100 ms), high reliability ($>$ 90%), and precise positioning (50 – 1000 m) underpin advanced scenarios.

To meet the requirements of diverse use cases in vehicular networks, 3GPP Rel 16 specifies three cast types for NR V2X, providing physical layer support for unicast, groupcast, and broadcast transmissions in the SL, which distinguishes it from LTE V2X that supports only broadcast transmissions. Unicast enables direct communication between a pair of UEs, while broadcast involves a single transmitter sending messages to be received by all UEs within its radio transmission range. Groupcast, or multicast, allows a transmitter UE to send messages to a specific set of receivers meeting required conditions. In addition, NR V2X incorporates a hybrid automatic repeat request (HARQ) procedure tailored for both unicast and groupcast messages.

In NR V2X SL communications, the data transmission process involves the organization of data into transport blocks (TBs), each accompanied by sidelink control information (SCI) carried in a physical SL shared channel (PSSCH) [42]. Different from LTE, NR V2X employs a two-stage transmission of SCIs, with the 1st-stage sent on the physical SL control channel and the 2nd-stage on the corresponding PSSCH. This dual-stage approach allows for flexible SCI content, supporting unicast, groupcast, and broadcast transmissions, providing a significant improvement over the limited broadcast capability. The split SCI design allows non-receiving UEs to decode only the 1st-stage information for resource allocation, while the 2nd-stage SCI provides detailed instructions for actual receivers. In addition, two HARQ feedback options include designated UEs within a specific distance of the transmitter and feedback from all receiving UEs, utilizing the physical SL feedback channel (PSFCH) in response to PSSCH transmissions. Additionally, for unicast scenarios, NR V2X supports CSI reporting, where the transmitting UE emits CSI-RSs for the receiving UE to measure and report back CSI via the PSSCH. Furthermore, NR V2X incorporates synchronization support for SL communications. A UE acting as a synchronization reference, called SR-UE, transmits synchronization information on the SL synchronization signal block (S-SSB), comprising the physical SL broadcast channel (PSBCH), SL primary synchronization signal (S-PSS), and SL secondary synchronization signal (S-SSS) [148]. Synchronizing to the SR-UE allows nearby UEs, potentially outside network or GNSS coverage, to establish an SL communication with both the SR-UE and each other. Notably, the SR-UE does not necessarily transmit data, emphasizing its role as a synchronization reference point.

LTE V2X typically operates in the LTE spectrum, while NR V2X operates in the 5G spectrum. NR allows flexible and efficient spectrum usage with scalable bandwidth and different modes of resource allocation. NR-V2X can support wider bandwidths, enabling higher data rates and better spectral ef-

ficiency than LTE V2X. NR V2X communication embraces flexibility and efficiency through BWPs [39]. These customizable portions within the broader carrier bandwidth allow customizing resource allocation to diverse V2X services and device capabilities [149, 150]. By supporting smaller BWPs, low-end UEs with limited power budgets can reduce their operational burden. Meanwhile, BWPs enable targeted allocation of resources based on specific service requirements, minimizing overhead and spectral waste. The BWP flexibility can simultaneously support various V2X services with differing bandwidth and numerology needs, maximizing carrier utilization.

In addition to BWPs, V2X communications are improved significantly through innovation in different modes of resource allocation. Two modes for NR V2X SL communications for sub-channel selection are proposed in Rel 16. *Mode 1* in NR V2X SL communication prioritizes predictability and network control by employing an allocation mechanism controlled by the gNB. This approach inherits dynamic grant scheduling utilized in LTE V2X *Mode 3* but replaces the semi-persistent scheduling with a more flexible configured grant scheduling [150]. Based on the dynamic grant, *Mode 1* UEs become highly communicative, requesting resources from the gNB for each TB, including retransmissions. This constant communication involves UEs sending scheduling requests through the PUCCH, to which the gNB responds with DCI on the PDCCH. To further enhance coordination, UEs inform each other about their allocated resources for TB and potential retransmissions using the 1st-stage SCI. It is seen that *Mode 1* allocation mechanism offers several advantages. The network awareness of gNBs enables optimized channel selection, scheduling, and power control, maximizing resource utilization and minimizing interference. By centrally managing resource allocation, the gNB ensures coordinated transmissions and collision avoidance, which are crucial for reliable communications, especially for safety-critical V2X applications. Additionally, the gNB can prioritize resource allocation based on UE needs and application requirements, promoting fairness and guaranteed QoS. However, the centralized approach of *Mode 1* also has drawbacks. Reliance on the gNB introduces a single point of failure. Additionally, centralized decision-making can increase latency, especially as network density grows. Moreover, the constant communication between UEs and the gNB can add to network overhead.

Mode 2, characterized by autonomous allocation, stands out for its distributed and self-organizing approach to resource management. Compared to the centralized control of *Mode 1*, UEs in *Mode 2* independently select communication channels based on local sensing and information sharing. This autonomy offers several key features. Its scalability and robustness are noteworthy, as the distributed nature inherently scales well with increasing network density. Individual UE decision eliminates dependence on a central gNB, making the system resilient to gNB failures. Secondly, *Mode 2* ensures low latency and adaptability crucial for real-time V2X applications. Specifically, direct channel sensing and localized resource selection enable faster allocation decisions, allowing UEs to adapt to dynamic channel conditions and traffic conditions, yielding a more interactive network. Moreover, *Mode 2* promotes efficient spectrum utilization by employing distributed allocation algorithms that can achieve higher spectrum efficiency than centralized schemes. UEs can exploit spatial reuse opportunities by dynamically avoiding occupied channels, thereby reducing interference and maximizing network capacity. However, the lack of centralized control in *Mode 2* can lead to potential fairness issues and higher interference risks due to competing UEs.

Two resource allocation modes presented in Rel 16 have a trade-off between decentralized agility and centralized efficiency. *Mode 1* prioritizes low latency and localized traffic adaptation through UE-driven scheduling, potentially incurring collisions and demanding complex UE capabilities. Conversely, *Mode 2* leverages the comprehensive view for optimized resource usage and collision avoidance, potentially introducing scheduling delays and necessitating robust network infrastructure. Choosing the optimal mode hinges on the specific V2X scenario, with latency-critical safety applications favoring *Mode 1*, while large-scale deployments with predictable traffic patterns benefit from coordinated orchestration in *Mode 2*. Ultimately, Rel 16 empowers V2X designs to select the suitable mode for the practical requirements of systems, ensuring efficient and reliable communication across diverse vehicular interactions.

HARQ is an extended feature for SL communication in NR that is provided by the MAC layer. the HARQ procedure enhances the reliability of TB transmissions through a retransmission process. Specifically, the receiver can request a HARQ retransmission if forward error correction and error detection codes fail to correct the entire transmission error. A HARQ response has different forms, such as an acknowledgment (ACK) in case of successful reception, a negative acknowledgment (NACK) if the reception is unsuccessful, or no response if the control information linked to the transmission is not successfully received within a preset timeframe. In unicast communications, the system employs ACK/NACK feedback for SL HARQ, where the receiver responds with ACK upon successful TB decoding or NACK if decoding fails after the 1st-stage SCI. For groupcast, two feedback options exist. We note that the concepts of two options mentioned in this section have no connection to those, i.e., *Options 1-8*, mentioned in Section 2.1. In *Option 1*, a receiver transmits NACK only if it fails to decode the TB and its distance to the transmitter is within the required communication range specified in the 2nd-stage SCI. This option, called NACK-only feedback, captures scenarios where feedback is unnecessary due to successful TB decoding or if the receiver is beyond the communication range. On the other hand, *Option 2* supports ACK/NACK feedback from all receivers in groupcast, indicating the success of TB decoding. In the context of groupcast, the transmitter specifies in the 2nd-stage SCI whether to use NACK-only feedback (*Option 1*) or ACK/NACK feedback (*Option 2*). NACK-only feedback is useful for groupcast services with less relevant information for receiving UEs beyond the communication range, such as in the extended sensors use case. For groupcast *Option 1*, the zone ID of a transmitter in the 2nd-stage SCI helps the receiving UE determine the Tx-Rx distance, enhancing communication reliability by considering the spatial relationship between the

transmitter and receivers. This innovation promotes a more efficient and adaptive communication framework, ensuring that feedback mechanisms are adjusted to the specific needs of each scenario, ultimately optimizing the NR V2X communication protocol.

In NR V2X SL communication, two HARQ feedback options offer distinct trade-offs between resource efficiency and feedback precision, as shown in Table 11. *Option 1* prioritizes resource conservation by using a shared channel for receiving UEs to transmit NACK-only feedback, indicating decoding failures within the specified range without individual UE identification. While minimizing resource consumption, this approach limits the ability of transmitters to customize retransmissions, potentially leading to unnecessary rebroadcasts. In contrast, *Option 2* allocates a dedicated resource for ACK/NACK feedback to each receiver, allowing fine-grained identification of successful and failed decodings. This precision enables targeted retransmissions, enhancing overall reliability but with a higher resource cost. The choice between options depends on the specific requirements of the V2X scenario. *Option 1* suits resource-constrained environments where basic reachability information suffices, while *Option 2* excels in scenarios requiring high reliability and efficient data delivery, albeit with increased resource demands.

3.7. NR for Unlicensed Spectrum

5G NR-U is a crucial mode of operation within 3GPP Rel 16 [4]. In addition to the NR FRs utilizing licensed bands defined in 3GPP Rel 15, NR-U provides the necessary technology for cellular operations to seamlessly integrate unlicensed spectrum into 5G networks. NR-U enables both UL and DL operations in unlicensed bands, supporting 5G features such as wideband carriers, flexible numerologies, dynamic TDD, beamforming, dynamic scheduling, HARQ timing, etc. Based on how the carriers are used for the UP and CP, three deployment modes within NR-U can be classified as CA, dual connectivity, and standalone [151]. Specifically, NR-U CA utilizes unlicensed spectrum to increase only the downstream UP capacity, whereas NR-U dual connectivity allows both upstream and downstream UP traffic using unlicensed spectrum. We note that both NR-U CA and NR-U dual connectivity modes define the CP traffic to be transported over the licensed spectrum. In the NR-U standalone mode, 3GPP Rel 16 first defines the operation that relies solely on unlicensed spectrum for both CP and UP traffic, i.e., without assistance from a carrier in licensed spectrum.

The deployment of 5G NR-U starts with the unlicensed 5 GHz band for the NR-U CA and NR-U dual connectivity modes. As new 6 GHz unlicensed spectrum becomes available in the USA (5925 – 7125 MHz) and the EU (5925 – 6425 MHz), with plans for availability in other countries soon, along with unlicensed mmWave spectrum (57 – 71 GHz), the 5G NR-U standard is poised to support these bands next, especially for NR-U standalone [152]. It is important to note that within NR-U standalone, the 6 GHz unlicensed spectrum is first defined in 3GPP Rel 16, whereas the unlicensed mmWave spectrum is later defined in 3GPP Rel 17. NR-U standalone eliminates any reliance on licensed network operators and is available for deployment by private enterprises, managed service providers, or network systems integrators. This allows private 5G implementations, which support emerging consumer and Industry 4.0 applications, to access secure, low-latency, reliable, and high-bandwidth connectivity for densely populated endpoints.

The primary concern of NR-U operation is to enable high utilization of 5 GHz bands while ensuring harmonious coexistence with Wi-Fi technologies. In association with NR-U, previous generations of Wi-Fi devices (i.e., IEEE 802.11a/b/g/n/ac) continue to operate in 5 GHz unlicensed bands, whereas IEEE 802.11ax is first standardized to allow unlicensed Wi-Fi operations in 6 GHz bands. We note that NR-U can employ the 5G NR PHY to maintain the 5G evolutionary benefits; however, MAC necessitates modification in NR-U to align with Wi-Fi technologies, where the contention-based LBT mechanism is perceived as the baseline for 6 GHz operations. There are two LBT categories: frame-based equipment (FBE) and load-based equipment (LBE). The mechanism of FBE and LBE is briefly summarized below.

As illustrated in Fig. 20(a), because NR-U devices in the FBE LBT mode perform channel sensing at fixed frame periods (FFPs), they must wait until the next frame interval for contention, even when the payload transmission can finish before the next frame boundary. Nevertheless, the NR-U work item [153] indicates that FBE LBT is efficient to NR-U only in environments where the absence of Wi-Fi networks is guaranteed. Therefore, this can limit the performance of FBE LBT associated with NR-U in general, which yields open challenges in designing synchronized access for FBE LBT. Muhammad *et al.* [154] investigated FBE LBT operating in dense deployment scenarios, evaluating channel utilization, collision probability, and channel access delay. They also presented a case study involving intensive care unit hospital environments enabled with 5G NR-U, where the aim was to emphasize how the choice of channel access parameters could influence the wireless coexistence of medical devices and characterize diverse risk profiles when operating in 5G NR-U.

Unlike the FBE LBT strategy, LBE LBT performs channel sensing at any instant in time, which allows the contention of NR-U devices to be experienced sequentially whenever the channel becomes idle, as depicted in Fig. 20(b). It is important to note that LBE LBT, used for NR-U devices, employs the same mechanism as carrier sense multiple access with collision avoidance, which uses an exponential back-off procedure implemented for Wi-Fi devices. The details of carrier sense multiple access with collision avoidance can be found in [155]. Compared to FBE LBT, which senses the channel during FFPs and consumes less energy than LBE but may not be suitable for heavy traffic patterns [154], LBE LBT involves continuous channel monitoring. In other words, devices in the LBE strategy transmit immediately when the channel is free, but they have to wait and retry if it is occupied to avoid collisions. Therefore, LBE is more efficient for bursty traffic but can be more complex and consume more energy. Based on the LBE LBT strategy associated with NR-U specified in 3GPP Rel 16, several analytical and experimental works can be reviewed as follows.

Table 11: Trade-offs of two NR V2X HARQ feedback options

	***Option 1* (NACK-only)**	***Option 2* (ACK/NACK)**
Resource Usage	Low	High
Target-specific retransmissions	No	Yes
Reliability	Lower	Higher
Advantages	Resource-constrained scenarios, basic reachability information	High-reliability scenarios, efficient data delivery
Disadvantages	Cannot identify specific failing UEs, potential unnecessary rebroadcasts	Higher resource consumption

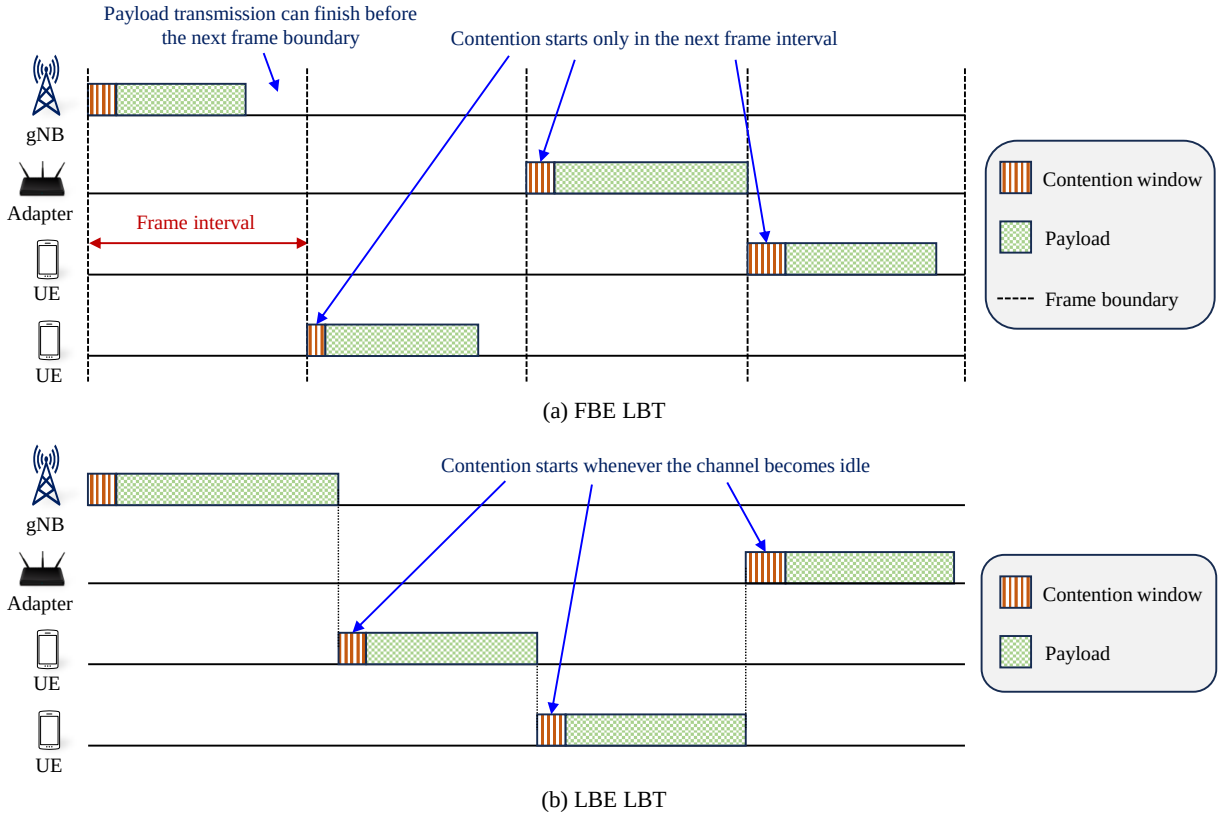


Figure 20: Illustration of LBT with (a) FBE and (b) LBE.

In [156], the impact of NR-U settings with flexible numerology and mini-slot transmissions was investigated in the context of DL channel access, considering the coexistence of field-programmable gate array-based license-assisted access, NR-U, and Wi-Fi prototypes. The OTA experimental results have demonstrated the bandwidth utilization efficiency and fairness of resource sharing among proposed prototypes in some cases. However, fairness is not guaranteed in general, which opens challenges in designing novel mechanisms that improve resource sharing between contention-based and scheduled technologies. Wang *et al.* [157] examined NR-U and Wi-Fi coexistence using the duty cycle strategy, where the bandwidth and transmission opportunity allocation were jointly optimized to maximize system throughput under fairness constraints. Numerical results have confirmed throughput improvements compared to certain benchmarks, validating the dynamic solution stability across various fairness thresholds and channel conditions. Subsequently, the NR-U type B multichannel access procedure for an NR-U and Wi-Fi coexistence system was introduced in [158]. The proposed framework focused on two unlicensed channels: a primary channel and a secondary channel. The system throughput and packet delay on each channel were analyzed by constructing Markov models. It is observed that increasing the payload size leads to throughput improvements for the Wi-Fi system on the secondary and for both the NR-U and Wi-Fi systems on the primary channel. However, it reduces the throughput of the NR-U system on the secondary channel. Furthermore, Loginov *et al.* [159, 160] proposed two novel collision resolution LBT schemes for NR-U and Wi-Fi coexistence. In the former strategy [159], an NR-U device randomly stopped sending the reservation signal to listen to the channel and then detected and resolved collisions. In contrast, in the latter strategy [160], a reservation signal was continuously sent until the next frame interval to enhance both channel resource efficiency and resource-sharing fairness. Their simulation results have revealed that both the proposed frameworks offer simultaneous throughput improvements for both NR-U and Wi-Fi systems.

In FBE, channel sensing occurs at fixed intervals, constrain-

ing the transmitter's access attempts when the channel is busy during a channel clear access event. On the other hand, LBE allows continuous channel sensing whenever data is present. As a result, low-priority LBE transmitters have a lower average access time compared to their FBE counterparts, even if they have an equal chance of encountering a busy channel. When low-priority LBE transmitters with packets that can tolerate longer latency requirements are being used, they can switch to FBE mode with an extended channel access time. This helps reduce the frequency of channel-sensing operations and lessen the workload on the gNB. This transition is particularly beneficial for energy-saving purposes in devices with limited power, even when dealing with high-priority URLLC packets with a lower data rate. The authors in [161] presented effective transitions between the two LBT mechanisms based on the channel access time and transmission probability derived from the Markov chain model. The experimental results in [161] clarify the extensive benefits of the proposed dynamic switch between LBE and FBE in reducing channel access time for high-priority data, such as URLLC, and improving energy efficiency for low-priority data in eMBB.

3.8. Summary and Discussion

The first step in 5G evolution, 3GPP Rel 16, studies several significant NR and LTE enhancements for 5G use scenarios. Those improvements mentioned in Rel 16 can be extensions of the existing work in Rel 15, whereas others are new features addressing new deployment scenarios. The novel technologies in 3GPP Rel 16 are summarized as follows.

As discussed in Section 3.1, the first key technical aspect is the enhancement of NR NOMA to accommodate numerous UEs without excessive cross-user interference [61]. In addition, it allows multiple users to transmit on the same subcarrier simultaneously but with different power levels to ensure decoding success. NR NOMA can be implemented in various mechanisms, as described in Fig. 8, with distinct key features. Grant-based NOMA ensures quality of service through pre-assigned resources but suffers from limited scalability and high complexity. RACH-based grant-free NOMA improves scalability by eliminating grant signaling but requires careful design for fairness and dynamic access. RACH-less grant-free NOMA maximizes efficiency with minimal signaling and complexity but presents challenges in collision avoidance and ensuring fair access for weaker UEs. In addition, several companies presented numerous compelling NOMA schemes, as listed in Table 6. Those approaches can be classified into three main categories: scrambling- interleaving-, and spreading-based NOMA schemes. Scrambling-based NOMA requires a low complexity; however, they are susceptible to interference. Interleaving-based methods combat channel variations and reduce latency. Nevertheless, it has the trade-off of supplementary overhead. On the other hand, while spreading-based schemes present notable advantages in terms of enhanced capacity gains and interference alleviation, they require higher complexity due to spreading sequences.

One of the most notable innovations addressed in the 3GPP Rel 16 is the IAB technology, which provides an alternative to the conventional fiber backhaul by extending NR to support wireless backhaul, as discussed in Section 3.2. The information from the central donor can be distributed to the neighbor nodes via multiple wireless links; thus, the IAB simplifies the deployment of small cells and resolves an overloading situation. The high-level procedure of IAB is based on the CU–DU split architecture, as described in Fig. 11.

For many years, the GNSS has successfully accomplished UE positioning. However, this technology is generally limited to outdoor scenarios with satellite visibility. For 3GPP Rel 16, several UE positioning techniques for indoor applications are reviewed in Section 3.3. Specifically, RAT-dependent techniques are introduced to meet the 5G positioning error requirements [110]. They can be classified as time-based (i.e., TDoA and RTT) and angle-based solutions (i.e., DoA). Both timing-based and angle-based positioning techniques offer distinct advantages and drawbacks, impacting their suitability for various scenarios. Timing methods offer wide applicability and infrastructure simplicity but suffer from multipath problems and synchronization dependence. Conversely, angle-based techniques boast superior precision and dense-environment resilience yet entail higher complexity and line-of-sight limitations. Combining both timing and angle-based techniques leverages their positioning performance while mitigating their weaknesses, as shown in existing studies in Table 8. In addition, DL and UL localizing procedures are supported by two new RSs, including DL PRS and UL SRS. Hybrid approaches aided by those RSs can achieve the positioning accuracy requirements, as shown in Table 7, and robustness in diverse environments [133, 134].

Low latency control has been a KPI for user experience since the early days of 5G communication. The two-step RACH has been specified in 3GPP Rel 16 to reduce the time for an initial UE–BS connection establishment. The mechanism of the two-step RACH is compared with that of the four-step RACH in Section 3.4. The two-step RACH can be implemented in a NOMA-based grant-free transmission to simultaneously reduce the signaling overhead and synchronize user signals. Moreover, the two-step procedure requires fewer processing steps than its counterpart; therefore, it has the advantage of power saving in a scenario where the UE data are transmitted intermittently.

The energy-efficiency techniques discussed in Section 3.5 have been studied and developed to improve the battery lifetime of terminal devices. The DRX and BWP adaptations are two conventional techniques proposed in 3GPP Rel 15 to enable energy-efficient mobile communication. Based on those two inventions, several enhancements and extensions were introduced in Rel 16, including the dynamic antenna adaptation, WUS indication, PCell signaling-based SCell dormant state transition, and cross-slot scheduling. These methods can be simultaneously implemented based on various use cases. For example, using DRX combined with the PDCCH optimization schemes (i.e., WUS or cross-slot scheduling) can reduce power consumption by up to 20% [162]. While those UE-level power-saving techniques in 3GPP Rel 16 offer significant benefits, they also involve trade-offs, as indicated in Table 10. The effectiveness of those methods relies on context-aware selection, where the choice of technique dynamically adapts to traffic

load, UE mobility, and channel conditions. For instance, prioritizing DRX and SCell dormancy during low-traffic scenarios minimizes unnecessary radio activity, while BWP management and MIMO layer reduction efficiently handle high-traffic demands.

The development of V2X communication is demonstrated through technical reports in Rel 16, where NR V2X technology plays a significant role [60, 147]. Compared to its predecessor, LTE V2X, presented in Rel 15, NR V2X shows significant advancements across KPIs. Specifically, as discussed in Section 3.6, its reduced latency enables real-time data exchange and unlocking applications, including collision avoidance and cooperative driving, which were previously unattainable with LTE V2X technologies. Furthermore, higher bandwidths in NR V2X communications accommodate the requirements of data-intensive transmissions, such as high-definition maps [163] and real-time video streaming [164]. Reliability also benefits from diverse modes of resource allocation, mitigating the reliance on cellular networks. These advancements pave the way for numerous novel V2X applications, from platooning and remote driving to real-time traffic optimization and immersive in-vehicle experiences.

Finally, Section 3.7 delves into the significance of 5G NR-U specified in 3GPP Rel 16. Based on the utilization of carriers for the UP and CP, NR-U classifies deployment modes as CA, dual connectivity, and standalone. The primary unlicensed bands designated for 5G NR-U operation are situated at 5 GHz and 6 GHz. The former (5 GHz) is preferable for NR-U CA and NR-U dual connectivity, whereas the latter (6 GHz) is favored for NR-U standalone. Due to the distinctive attributes of NR-U standalone, which operates exclusively on the unlicensed spectrum for both CP and UP traffic, it emerges as a suitable choice for private 5G implementations. These implementations boast secure access, low latency, high reliability, and robust high-bandwidth connectivity, especially pertinent for densely populated endpoints. Capitalizing on these advantages, NR-U standalone has been expanded into the unlicensed mmWave spectrum in 3GPP Rel 17. Notably, the harmonious coexistence between 5G NR-U and Wi-Fi technologies is a principal consideration. In the realm of NR-U MAC, a contention-based LBT mechanism is employed, comprising two categories: FBE LBT and LBE LBT. Each category employs distinct sensing strategies and addresses specific challenges. FBE LBT proves efficient for NR-U in environments where the absence of Wi-Fi networks is assured. In contrast, LBE LBT, reminiscent of Wi-Fi technology's carrier sense multiple access with collision avoidance, allows contention at any time instant, potentially enhancing efficiency in coexistence environments.

Building upon technologies in Rel 15, Rel 16 of the 3GPP standard emerged as a crucial stepping-stone in the evolution of 5G communications. While Rel 15 introduced core concepts, such as URLLC and massive MIMO, Rel 16 significantly enhanced their performance and flexibility, paving the way for the advanced features of Rel 17. In addition, IAB, introduced in Rel 16, allows for consistent merging of cellular and Wi-Fi networks, optimizing resource allocation and improving user experience. This provides cost-effective network deployments and wider 5G coverage, particularly in dense urban environments and indoor settings. Furthermore, Rel 16 placed significant emphasis on latency reduction. Enhanced positioning and latency reduction features were introduced, enabling near-real-time responsiveness for applications such as augmented reality, virtual reality, and industrial automation. These improvements laid the groundwork for the even more stringent ultra-low latency demands addressed in Rel 17 through reduced capability NR devices. Another key contribution of Rel 16 was the formalization of NR V2X communications. This technology ensures safer and more efficient transportation systems by enabling communications between vehicles, infrastructure, and pedestrians. While further enhancements are expected in Rel 17, Rel 16 laid the crucial groundwork for V2X integration into 5G Advanced networks.

Table 12: Representative RedCap use cases in 3GPP Rel 17.

	Industrial sensors	**Surveillance**	**Wearables**
Bit rate	2 Mb/s	2–4 Mb/s for economic video; 7.5–25 Mb/s for high-end video	UL: 2–5 Mb/s (up to 50 Mb/s); DL: 5–50 Mb/s (up to 150 Mb/s)
Latency	100 ms (5–10 ms for safety sensors)	500 ms	-
Reliability	99.99%	99–99.9%	-
Battery life	At least a few years	-	At least a few days and up to 1–2 weeks
Traffic type	UL heavy	UL heavy	-
Mobility	Stationary	Stationary	Non-stationary

4. Release 17

3GPP Rel 17 is the third major iteration of the 5G standard [5]. Rel 7 significantly enhances the core functionalities of 5G technology, including coverage, mobility management, power efficiency, and reliability. Additionally, it expands the applicability of 5G by enabling novel use cases, deployment scenarios, and network configurations. In this section, we present five crucial aspects of this release: reduced capability (RedCap) NR devices (RedCap), non-terrestrial network (NTN), NR multicast-broadcast services (MBS), edge computing, and RAN slicing.

4.1. RedCap NR Devices

The ongoing expansion of 5G ecosystems faces challenges in broadening device connectivity and catering to diverse use cases. In response, a key development involves the convergence of eMBB, mMTC, and URLLC with emerging IoT applications across vertical industries. To address these challenges, 3GPP introduced RedCap NR devices in Rel 17 [165, 166]. This initiative focuses on enabling new device categories, including industrial wireless sensors, video surveillance systems, and wearables. Table 12 summarizes the specific specifications

Table 13: Comparison of the baseline Rel 15 baseline UE and RedCap NR UE.

	FR1		FR2	
	Baseline UE	**RedCap UE**	**Baseline UE**	**RedCap UE**
Number of receiver branches	2 or 4	1	2	2
UE bandwidth	100 MHz	20 MHz	200 MHz	100 MHz
Number of MIMO layers	2 or 4	1	2	1
Maximum modulation order	64 QAM	64 QAM	64 QAM	64 QAM
Duplex operation	TDD or full-duplex FDD	TDD or half-duplex FDD	TDD	TDD
Cost reduction	0%	–65%	0%	–50%

Table 14: Types of NTN platforms.

Platform	**Altitude Range (km)**	**Orbit**	**Beam Footprint Size (km)**
GEO	35786	Fixed position	200–3500
MEO	7000–25000	Circular around Earth	100–1000
LEO	300–1500	Circular around Earth	100–1000
UAS	8–50 (20 for HAP)	Fixed position	5–200

for these use cases. RedCap recognizes the diversity in requirements across these use cases and is designed to navigate the trade-off between cost/complexity, battery life, and varying performance demands. Consequently, it emerges as a well-suited solution for supporting mid-tier IoT deployments. Beyond fulfilling specific use case requirements, RedCap NR UE aims to significantly reduce cost/complexity and physical size compared to standard 5G NR UEs (e.g., eMBB and URLLC devices) while extending battery life. RedCap NR UE supports all FR1 and FR2 bands for both FDD and TDD operations. RedCap incorporates various features to meet diverse requirements, including complexity reduction and enhanced energy-efficiency techniques for UEs.

Regarding UE complexity reduction, the Rel 17 RedCap standard has implemented various features, including a limited number of receiver/transmitter branches, reduced UE bandwidth, half-duplex FDD operation, maximum number of MIMO layers relaxation, and maximum modulation order relaxation. To compare the capabilities of a baseline Rel 15 UE with the RedCap NR UE, Table 13 presents the differences. It shows that the decreased capabilities can result in approximately 65% and 50% reduction in the UE bill-of-materials expense for FR1 and FR2, respectively [166]. This implies significant cost savings without compromising performance, specification, and coexistence effects. Furthermore, the ability to handle a single antenna branch facilitates smaller device form factors, which is particularly important for wearables [167]. The study in [168] determined how these complexity reduction strategies may affect coverage. The findings suggest that RedCap NR channels require a coverage recovery of less than 1 dB or can be compensated for using reduced data rates.

Regarding UE energy savings, two specified techniques are extended DRX in RRC idle/inactive states and relaxed radio resource management (RRM) measurement for adjacent cells. The extended DRX enables the UE to stay in a low-power state for an extended period, with DRX cycles of up to 10485.76 s (approximately 3 h) in the RRC idle state and 10.24 s in the RRC inactive state. Compared to the legacy DRX sleep periods of 1.28 s or 2.56 s, this improvement provides significant battery life benefits for RedCap NR devices, such as industrial sensors. In the RRC idle/inactive state, the UE regularly conducts RRM measurements to facilitate smooth mobility management. These measurements are based on RSRP data from the current and adjacent cells. However, when the RSRP of the current cell reaches a certain threshold, the UE is not required to conduct RRM measurements for adjacent cells to save energy. Additionally, if the UE is stationary or not located at the cell border, the relaxation of RRM measurement for adjacent cells can conserve UE power. In the RRC connected state, the network can further establish a stationary condition based on RSRP for the UE. This allows the UE to report to the network when this condition is met, resulting in further power savings.

4.2. Non-Terrestrial Network

NTN has gained significant attention in Rel 17 due to its support for ubiquitous connectivity and coverage, separating the technology into two facets: NR NTN and IoT NTN [169, 11]. NR NTN leverages the power of 5G NR technology to provide broadband services via satellite constellations in FR1. IoT NTN, on the other hand, focuses on connecting massive numbers of low-power, low-complexity devices using existing narrow-band technologies such as narrow-band IoT (NB-IoT) and enhanced machine-type communication (eMTC) over satellite links. Depending on the type of NTN platform, various deployment options are available. These platforms can be classified into two primary groups: spaceborne and airborne. The classification is based on altitude, orbit type, and beam footprint size, as summarized in Table 14.

- Spaceborne vehicles: Recognizing the potential for shared growth, the satellite industry has been a driving force behind integrating satellite or spaceborne platforms into the 5G ecosystem through the 3GPP initiative. Spaceborne platforms can be categorized into three types: low Earth orbit (LEO), medium Earth Orbit (MEO), and geosynchronous Earth orbit (GEO) satellites. GEO satellites remain stationary in space relative to terrestrial observers. LEO and MEO satellites, a.k.a. non-GEO satellites, have orbital periods ranging from 1.5 to 10 hours, shorter than the Earth's rotation time.

- Airborne vehicles: They encompass two categories, i.e., uncrewed aircraft systems positioned within altitude ranges of 8 to 50 km and HAPs navigating at heights of 20 km. Comparable to geostationary satellites, uncrewed

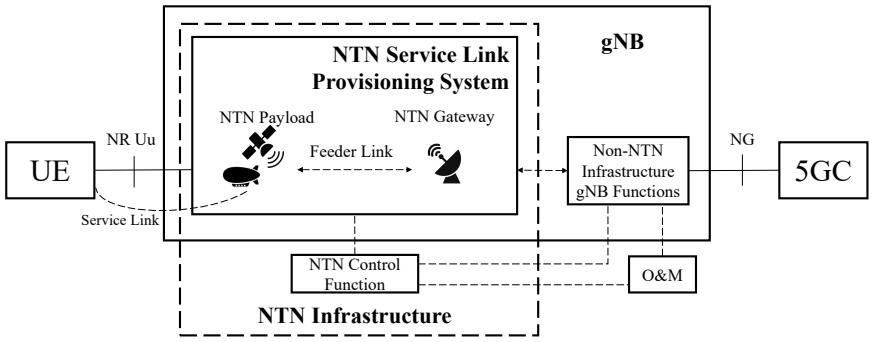


Figure 21: NTN architecture with transparent payload in Rel 17.

aircraft systems can remain stationary relative to a ground reference point. The beam footprint size of uncrewed aircraft systems spans from 5 to 200 km.

In Rel 17, the 5G NR-based NTN study focused on transparent (non-regenerative) payloads in LEO and GEO network scenarios. These scenarios involved Earth-fixed tracking regions, FDD systems, and GNSS-capable UEs. GNSS-equipped UEs leverage their positioning capabilities to estimate the relative speed and RTT to the satellite. This enables them to apply precise Doppler pre-compensation to guarantee accurate UL signal transmission on the desired frequency [170]. The NTN architecture with transparent payload is illustrated in Fig. 21. The NG interface links the 5GC network with a gNB, which is located on the ground behind the NTN gateway. The NTN gateway connects to the NTN payload through the feeder link. The NTN payload communicates with the UE over the service link through the NR Uu interface.

NTN, coupled with existing terrestrial infrastructure, could deliver cost-effective solutions for many use cases [171, 172]. NTN can provide continuous and ubiquitous wireless coverage for NB-IoT/eMTC devices. This enables broadband services in unserved or underserved areas and facilitates remote monitoring and control of diverse assets and infrastructure, ranging from transportation systems (ships, trains, trucks) to critical utilities (bridges, pipelines, railways) and environmental monitoring sensors. The inherent wide-area broadcast capabilities of NTN provide a foundation for novel mobile edge applications within the 5G ecosystem, exemplified by applications such as mobile gaming, where content needs to be readily available across geographically distributed edge locations. Offloading computational tasks to NTN can help terrestrial infrastructure handle growing demands [173]. Moreover, in a natural disaster or emergency that causes temporary network outages or destruction, NTN can serve as a reliable fallback to reestablish communication networks with minimal delay. In recent years, diverse industrial developments, such as SpaceX's Starlink [174], Amazon's Project Kuiper [175], and OneWeb [176], have unlocked the potential of LEO non-geostationary satellite constellations for global coverage. In addition, several field trials have assessed the performance of 5G in NTN configurations [177, 178, 179].

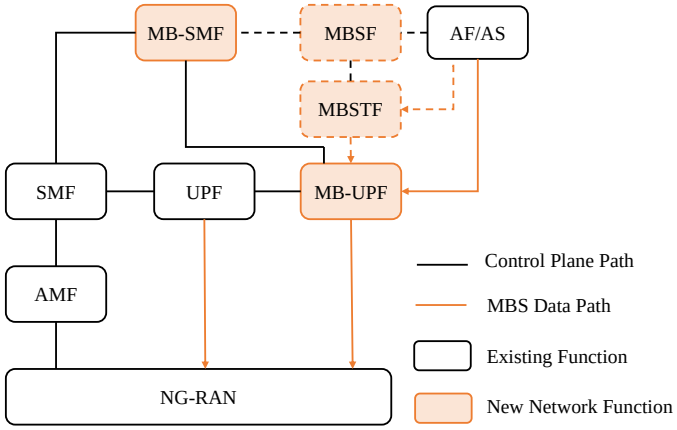


Figure 22: 5G architecture with MBS enhancements.

4.3. NR Multicast-Broadcast Services

With media streaming and increasing traffic, broadcast and multicast techniques offer a key solution for optimizing network resource efficiency. Rel 17 has made significant efforts to integrate MBS into the NR 5G system [180, 181]. MBS leverages the existing 5G NR and 5GC architecture with specific enhancements to simultaneously enable efficient content delivery to large groups of users. This represents a significant shift from the traditional unicast approach, where individual users receive independent data streams. The 5G MBS supports key use cases for public safety, OTA software updates, video delivery, connected vehicles, NTN, and other IoT applications.

As shown in Fig. 22, the 5G network architecture has been enhanced to support MBS while utilizing the existing 5G system [182]. The access and mobility management function is a CP interface with NG RAN to serve the UE. A session management function establishes a data channel between the UE and a UP function to enable the transmission of UE data. Several new network functionalities have been introduced to facilitate MBS. The multicast-broadcast UP function acts as an entry point to the 5G system and functions as a session anchor. The multicast-broadcast session management function manages MBS sessions and configures the multicast-broadcast UP function based on policy rules. The MBS function provides service-level capabilities for MBS session operations by connecting with the application function/application server and the

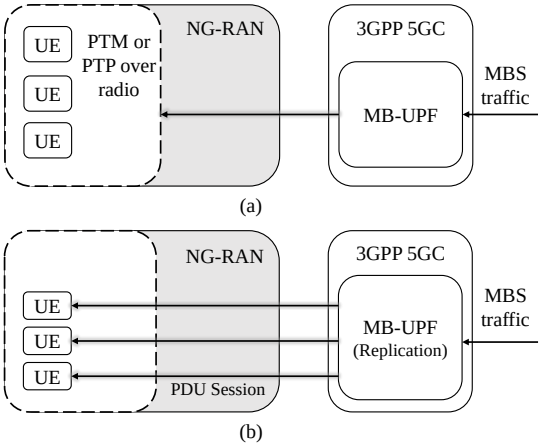


Figure 23: 5G MBS delivery mechanisms: (a) 5GC shared MBS traffic delivery, (b) 5GC individual MBS traffic delivery.

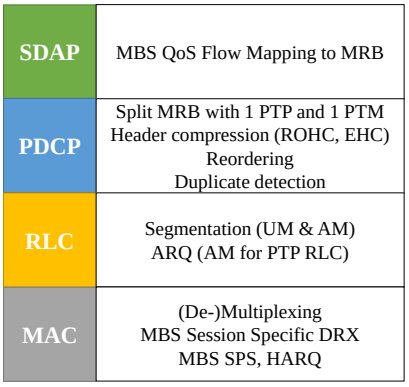


Figure 24: NR multicast-broadcast service protocol.

multicast-broadcast session management function. Lastly, the MBS transport function acts as a media anchor for MBS data traffic by offering generic packet transport functions to internet protocol multicast-enabled applications.

The MBS traffic is distributed to multiple UE from a single data source using two delivery mechanisms: 5GC shared and 5GC individual MBS traffic delivery, as shown in Fig. 23. The 5GC shared mechanism is used for multicasting and broadcasting, where copies of MBS packets are received and distributed by 5GC to the NR RAN node. This mechanism allows MBS packets to be sent via the radio between the NG RAN and UE through point-to-point (PTP) and point-to-multipoint (PTM). In the PTP scheme, each piece of UE receives distinct copies of MBS packets from the NR RAN node over the radio. In the PTM scheme, a group of UE receives copies of MBS packets from the NR RAN node over the radio. On the other hand, the 5GC individual MBS traffic delivery mechanism is exclusively for multicast services. In this mechanism, 5GC obtains copies of MBS packets and provides each UE with distinct copies of MBS packets through per-UE PDU sessions.

To ensure the QoS of multicast services, UEs must maintain the RRC connected state for proper radio resource configuration, including MBS radio bearer and PHY configurations. When cell edge users experience poor channel quality and require reliable transmission, the PTM transmission can be switched to PTP transmission. This change allows for utilizing an automatic repeat request (ARQ) operation, improving overall performance. If no multicast data is received, the multicast session can be disabled. To reactivate the multicast session, group paging is used to notify the UE to transition from the RRC idle/inactive state to the RRC connected state. On the other hand, broadcast services can serve all UEs within the coverage area irrespective of their RRC states. The radio resource configuration for broadcast mode is disseminated periodically via the dedicated MBS control channel. This allows UEs, regardless of their RRC state, to acquire the necessary settings from the MBS traffic channel and decode broadcast data. In addition, MBS delivers varying levels of data transmission reliability. HARQ enables swift physical layer retransmissions with fast feedback. At the same time, the network can dynamically choose between PTP or PTM retransmissions or even the robust ARQ for 100% delivery guarantees. Broadcast can leverage data bundling for enhanced reliability. Notably, multicast guarantees packet-level service continuity with lossless mobility by employing packet-level sequence number synchronization and allowing UEs to request missing packets during handover, ensuring a seamless experience.

Fig. 24 illustrates the architecture of the NR MBS protocol, which leverages and enhances the functionalities of unicast to support MBS. The protocol comprises four sub-layers: SDAP, PDCP, RLC, and MAC. The incoming QoS flows are mapped to multicast or broadcast service radio bearers using a service data adaptation based on the QoS policies. The PDCP executes packet reordering and links these packets to RLC entities. Header compression approaches (e.g., robust header compression and Ethernet header compression) can be used to reduce header overhead. The RLC segments the packets into smaller subpackets and is divided into PTM RLC and PTP RLC. The PTM RLC can deliver the same packet to multiple pieces of UE via a group radio network temporary identifier but only supports the unacknowledged mode of delivery due to the synchronization complexity. The PTP RLC employs a cell radio network temporary identifier and can deliver in the acknowledged mode. The MAC performs multiplexing/demultiplexing. The MBS session-specific DRX can be configured to conserve UE power for the UE that obtains MBS data. The MAC also manages MBS PHY techniques, including the HARQ operation and MBS semi-persistent scheduling.

4.4. Edge Computing

Edge computing was standardized by 3GPP and the European Telecommunications Standards Institute to facilitate time-sensitive tasks. This approach involves deploying communication infrastructure and computing resources closer to the UEs at the network edge [183]. Edge computing offers numerous advantages in enhancing 5G networks and enabling a wider range of innovative use cases. Some notable examples include real-time gaming, extended reality, connected vehicles, and industrial IoT applications [184].

In Rel 17, 3GPP aims to deliver native support of edge computing in 3GPP networks. Hence, it is necessary to enhance

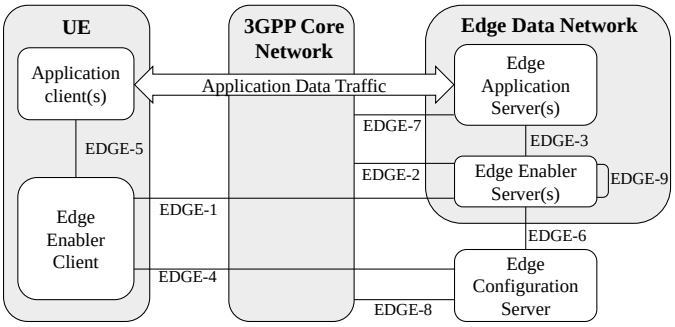


Figure 25: Architecture enabling edge applications.

the overall application layer architecture. The framework for enabling edge applications, depicted in Fig. 25, has been developed with various architectural principles, including flexible deployment, service differentiation (the ability to activate or disable edge computing functionalities), UE application portability, and edge application portability [185]. The application client can locate the optimal edge application server (EAS) by introducing the edge enabler client (EEC) and the edge enabler server (EES). Furthermore, the EEC can connect with the most suitable EES by utilizing the edge configuration server (ECS), which provides information regarding edge configurations. The EES, ECS, and EAS can communicate with the 3GPP core network. By leveraging the capabilities of the enabling layer, 3GPP networks can support the edge capabilities as follows.

- Service provisioning: The EEC–equipped UE can be assisted in locating and connecting to accessible edge data networks.
- Rich discovery: The EEC–equipped UE can determine the edges and EASs through on-demand configuration provisioning through the ECS and query support by the EES.
- Dynamic availability: The fluid nature of the system, including flexible deployments and UE movement, leads to dynamic changes in edge and EAS availability. Accordingly, UEs can subscribe to these dynamic changes, which allow them to adjust the offered services and maintain an optimal experience.
- Capability exposure: EES capabilities are made available to EAS as value-added services through application programming interfaces (APIs). Both EAS and EES can also leverage 3GPP network capabilities through northbound APIs provided by service capability exposure functions and network exposure functions.
- Support for service continuity: When UEs move across the network, the architecture intelligently migrates their application context to the most suitable server, whether on the edge or in the cloud. This seamless transfer guarantees uninterrupted service and preserves the user experience.

In Rel 17, several enhancements are described to assist applications in edge computing, i.e., EAS discovery, edge relocation, network exposure to EAS, and application function guidance to determine UE route-selection policy rules [186]. The EAS discovery function is defined to support PDU sessions with a session breakout connectivity model. It operates as a domain name system resolver for the UE, supplementing domain name system requests with UE mobility information. This approach allows the domain name system to resolve to EASs near the UE. When the EAS relocates due to UE mobility, the 5G system UP path can be reconfigured with the application function to maintain the optimized path and minimize the effect on the user experience. The application function and network-triggered edge relocation mechanisms are designed to accommodate various application requirements (e.g., packet loss and UP latency).

In addition, network exposure with low latency is specified to expose QoS monitoring results to EAS. With this UP function-based network exposure, the exposure path is shortened to accelerate fast application response to network changes. Finally, in the UE, the configuration of the UE route-selection policy rules can consider specific application server information. Thus, this enables the UE to establish PDU sessions for specific application servers flexibly, removing the requirement for deploying complex session breakout solutions. Some other important aspects are also discussed, including security [187], media processing [188], and management [189].

4.5. Radio Access Network Slicing

5G is envisioned as a multi-service network supporting various verticals with diverse performance and service requirements. The slicing concept, facilitating network virtualization techniques to create many logically separate networks, has emerged as an efficient way to serve all services on a common infrastructure [190]. RAN slicing splits RAN resources to produce various RAN slices, each designed to satisfy the requirements of a specific 5G service, i.e., eMBB, URLLC, or mMTC [191, 192, 193]. Mobile network operators consist of infrastructure providers, which own and control the physical resources (e.g., BSs, core network components, and radio resources), and mobile virtual network operators, which lease resources from providers to offer services to users. RAN resources are allocated based on service-level agreements (SLAs) and user needs. Hence, resource allocation is critical to RAN slicing, ensuring optimal network performance and service differentiation. Maule *et al.* [194] demonstrated through a single-tenant testbed with two slices (eMBB and mMTC) that dynamically adjusting slice configurations based on both SLAs and real-time traffic patterns can optimize RAN slicing performance. Albonda *et al.* [195] addressed resource allocation in a RAN slicing scenario with two service demands (V2X and eMBB), proposing a hybrid algorithm combining offline reinforcement learning for initial resource split and a low-complexity heuristic for dynamic adjustments.

In 3GPP, the identification of a network slice is defined by a single network-slice selection assistance information (S-NSSAI), including two parts: a slice/service type and a slice differentiator. The slice/service types describe unique identifiers for different network slice types, each optimized for specific use cases. Common standardized slice/service type val-

ues include 1 for eMBB, 2 for URLLC, 3 for massive IoT, 4 for V2X, and 5 for high-performance machine-type communications. The slice differentiator distinguishes several network slices of the same slice/service type. The UE enrolls in the 5G network for service delivery and establishes a 5G connection as a PDU session with a specific S-NSSAI. From the viewpoint of QoS management, one or multiple QoS forwarding treatments, a.k.a. 5G QoS flows, can be executed within each PDU session related to the given S-NSSAI as per the 5G QoS model.

3GPP Rel 17 investigated enhancements of RAN slicing for NR [196], such as enabling UE quick access to the cell serving the desired slice (e.g., the slice-aware cell reselection and slice-specific RACH configuration) and service continuity support for intra-radio access technology handover service interruptions. A new network slice as a group method is provided to avoid exposing S-NSSAI through the NR Uu interface for security and overhead concerns. This technique enables slice-aware cell reselection and slice-specific RACH configuration. Rather than the S-NSSAI, the data for the new network slice as a group method are broadcast in the system information. In the case of a slice resource shortage, the NG RAN assigns resources to the slice via multi-carrier resource sharing or resource repartitioning to ensure service continuity. The NR RAN can create dual connectivity or CA with different frequencies and overlapping coverage if the same slice is available for multicarrier resource sharing. Resource repartitioning enables a slice to employ resources from a shared or prioritized pool when resources are unavailable.

4.6. Summary and Discussion

3GPP continues 5G evolution in its Rel 17. In this section, we addressed the major new use cases and deployment scenarios, including RedCap, NTN, NR MBS, edge computing, and RAN slicing.

First, Section 4.1 introduces RedCap, which was first introduced in Rel 17 and marked a transformative step with its effective cost, energy efficiency, and compact design for UEs. RedCap fulfills the requirements of mid-range IoT use cases, including industrial sensors, video surveillance, and wearables. Further enhancements for RedCap are expected in the next 3GPP releases. The enhancements could further reduce RedCap NR UE cost/complexity/energy consumption and support emerging 5G applications such as smart city and eHealth coverage.

We review NTN in Rel 17 in Section 4.2. NTN breaks new ground in this release by integrating satellites seamlessly into the 3GPP ecosystem and forging a global standard for future satellite networks. It tackles coverage gaps and disruptions in underserved areas, boosts reliability through interconnectivity with diverse technologies, and strengthens network resilience against disasters. Rel 17 lays the foundation with NR NTN and IoT NTN standards, facilitating enhancements planned in the next 3GPP releases, such as enhanced coverage, improved mobility, and expanded spectrum support.

Section 4.3 presents NR MBS, which was first introduced in Rel 17. It has two delivery modes: multicast (for joined UEs in RRC connected state, targeting high QoS requirements) and broadcast (to all UEs in a service area, aiming for lower QoS requirements). MBS enhancements are expected in the next 3GPP releases, including enabling multicast reception in RRC inactive state for wider reach and lower UE power consumption, allowing UEs to receive unicast and broadcast transmissions simultaneously, and improving service provisioning in shared-network scenarios with multiple operators.

In Section 4.4, we review edge computing, which stands as a cornerstone for realizing the full potential of 5G applications, especially those demanding ultra-low latency. In Rel-17, the 5G system architecture has integrated the EAS discovery function, leveraging domain name system capabilities, to facilitate the discovery of edge applications near UEs. Future 3GPP releases aim to further enhance edge computing capabilities by streamlining the exposure of device-traffic-related information to EASs and optimizing the allocation and relocation of EASs among different UEs.

Finally, Section 4.5 outlines enhancements of RAN slicing for NR, a feature built on the existing 5G RAN that divides RAN resources into virtual slices tailored to specific 5G services. Some focus areas in the next 3GPP releases include UE-driven roaming partner selection based on supported slices, service continuity during SLA breaches, temporary slice support, and zero-touch network slice management.

It is worth noting that 3GPP Rel 17 was a testament to the resilience and ingenuity of the mobile industry in the face of a global pandemic. It marked the first entirely remote development process for a major release, and the results are impressive. Rel 17 paves the way for exciting new applications and use cases, laying the groundwork for the next chapter in mobile technology: 5G Advanced.

5. Release 18 (as Expected)

Recently, 3GPP officially announced the approval of the Rel 18 work package [197]. This is a significant milestone as it is the inaugural standard release of the 5G Advanced. Compared to previous release standards, 3GPP Rel 18 introduces substantial upgrades in system features [198, 199]. The 3GPP RAN Rel 18 package can be categorized into three main parts: eMBB evolution, non-eMBB evolution, and cross-functional evolution. This section discusses some noteworthy features of Rel 18, such as RAN intelligence, network energy savings, small data transmission, NR support for UAVs, low-power wake-up signal and receiver, and dynamic spectrum sharing (DSS).

5.1. RAN Intelligence

3GPP Rel 18 delves into normative work on AI/ML for RAN intelligence [200], focusing on advancing data collection and signaling capabilities to enable AI/ML-driven network energy conservation, load balancing, and mobility optimization. These enhancements seamlessly integrate with existing NG RAN interfaces and architecture, supporting both monolithic and split gNB configurations. AI/ML model training and inference can be flexibly deployed, either in the gNB (or gNB-CU for split

configuration) or in the operations, administration, and maintenance (OAM) function for training and the gNB (or gNB-CU) for inference. RAN intelligence is driven by predictive AI/ML abilities, which can be shared between gNBs via the Xn interface.

In addition, 3GPP Rel 18 initiated a study on the application of AI/ML for NR air interface [201]. The study aims to develop a standardized AI/ML framework, identify areas where these technologies can enhance NR air interface functions, define descriptions and characteristics for AI/ML models, and evaluate the performance, complexity, and potential specification impacts of these methodologies. Three use cases were prioritized for performance improvements: CSI feedback to optimize efficiency, accuracy, and predictive capabilities; beam management to enhance temporal/spatial beam prediction for accuracy and efficiency gains; and positioning accuracy across diverse scenarios, including non-line-of-sight situations. While the AI/ML model can be deployed at one side (either gNB or UE) for most use cases, the spatial-frequency domain CSI compression requires a two-sided model, which leverages an AI/ML-based encoder at the UE to efficiently compress CSI data, followed by its reconstruction at the gNB by a corresponding AI/ML decoder. Hence, a tight UE-gNB interaction is required in the two-sided AI/ML deployment model.

5.2. Network Energy Savings

Energy conservation is crucial for sustaining the environment, reducing environmental impacts, and minimizing operating costs for mobile network operators. Notably, 5G NR exhibits significant energy efficiency improvements over past generations. However, the dense deployment, extensive use of massive MIMO, wider bandwidths, and additional frequency bands in 5G networks can increase power consumption if appropriate energy-saving measures are not implemented. Hence, 3GPP approved a special study for network energy conservation in 5G NR in Rel 18 [202]. It focuses on developing and identifying an evaluation methodology, a BS network energy consumption model, and KPIs such as spectral efficiency, user throughput, latency, capacity, and UE power consumption. Additionally, techniques for energy conservation in gNB and UE regarding transmission and reception are explored and evaluated in specific deployment scenarios, including urban micro and macro settings with massive MIMO.

5.3. Small Data Transmission

Small data packets are commonly used in small IoT devices and wearables applications. When a new packet arrives, the UE changes from the RRC inactive state to the RRC connected state to transfer the data using four-step and two-step random-access techniques. The UE stays in the RRC connected state and performs additional procedures (e.g., measurement reporting) until it receives an RRC connection release message. These RRC state transitions can result in unnecessary control signaling overhead, especially when dealing with infrequent and power-sensitive devices. To address this issue, in 3GPP Rel 17, mobile-originated small data transmission (MO-SDT) is introduced [203]. In MO-SDT, the UE initiates the UL transmission while it remains in the RRC inactive state. 3GPP Rel 18 introduces mobile-terminated small data transmission (MT-SDT), allowing DL-triggered small data to be sent from the network to the UE with RRC inactive state [204]. The work item on MT-SDT for NR in Rel 18 aims to describe support for paging-triggered SDT. It includes two mechanisms for triggering MT-SDT, i.e., random-access-based SDT and configured grant-based SDT. It also specifies the MT-SDT process for receiving initial DL data and follow-up UL/DL data transmissions in the RRC inactive state.

5.4. Unmanned Aerial Vehicles

UAVs are increasingly gaining popularity in various applications [205]. For example, UAVs, acting as aerial BSs, can dynamically adjust their altitude and position, utilizing communication protocols to deliver Internet access to UEs. 3GPP standardized enhanced LTE support for UAVs in Rel 15 to Rel 17 [206]. However, compared to LTE, 5G NR offers a wider range of applications for UAVs with lower communication delays and higher data rates for services. Rel 18 introduced 5G NR support for UAVs [207], aligned NR solutions with the existing LTE UAV solutions as well as specified NR-specific enhancements. Several enhancements include UAV-related measurement reports (e.g., altitude threshold-based UE-triggered measurement reports and reporting of UAV spatial and kinematic data), UAV identification broadcast, UAV beamforming capability indication for UEs, and subscription-based UAV identification.

5.5. Low-Power Wake-Up Signal and Receiver

The study of the low-power wake-up signal (LP-WUS) and low-power wake-up receiver (LP-WUR) for NR was approved in 3GPP Rel 18 [208]. It is recognized as an iconic technology for UEs in the 5G Advanced. Controlling energy use is crucial for wearables and IoT devices, e.g., industrial sensors and controllers. The UE can save energy using DRX, which allows it to deactivate its transceivers for a DRX cycle when traffic data is not received [139]. However, the UE must wake up after each DRX cycle, resulting in energy waste. To resolve this issue, the UE can be outfitted with the LP-WUR that monitors the LP-WUS from the gNB to trigger the UE from idle/inactive to active. In Rel 18, the LP-WUS is not limited to using existing signals. Hence, investigations are conducted on LP-WUR architectures, LP-WUS designs, and procedures of the WUS to achieve significant advantages regarding energy efficiency, coverage availability, and latency impact over existing power-saving strategies.

5.6. Dynamic Spectrum Sharing

Effective management of the scarce spectrum resource is essential to maximize social benefits. To achieve this, 3GPP Rel 18 has introduced further improvements to NR for more adaptable and efficient spectrum utilization in 5G deployments, i.e., DSS [209], a.k.a. LTE-NR coexistence, which allows a BS to utilize a shared spectrum to give access to both LTE and NR UEs. This promotes spectrum transfer from LTE to NR,

enhancing NR spectrum efficiency, especially as fewer LTE devices will use the DSS carrier in the future. However, there are challenges in implementing DSS, particularly with the PDCCH. In DSS, the NR PDCCH and LTE PDCCH must share the first three OFDM symbols in a slot, with the restriction that symbols used by the NR PDCCH cannot be the same as those used by the LTE common reference signal. To address this issue and improve resource consumption and PDCCH capacity for DSS, Rel 18 provides support for the NR PDCCH reception in symbols with LTE common reference signal. In addition, the functionality of setting the UE with different LTE common reference signal rate-matching patterns in multiple TRPs is provided in a single TRP situation to reduce intercell interference.

5.7. Summary and Discussion

3GPP Rel 18 marks the start of 5G Advanced. As elaborated upon in Section 5.1, the inclusion of AI/ML in 5G Advanced is a critical milestone for the NG RAN and NR air interface. 3GPP Rel 19 anticipates undertaking normative work on integrating AI/ML into the NR air interface, informed by the outcomes of the Rel 18 study. Furthermore, novel use cases such as AI/ML-based mobility management will be investigated, while additional studies will delve into areas requiring further exploration, including testing methodologies for two-sided AI/ML models. Section 5.2 presents the special study for network energy savings. While minimizing radio network energy consumption is crucial, a holistic approach is vital to assess the trade-offs between energy savings and performance, encompassing KPIs, i.e., spectral efficiency, UE throughput, delay, capacity, and energy consumption. The work item in enhancements of the network energy savings is endorsed in Rel 19. Section 5.3 overviews the SDT feature, allowing data and/or signaling transmission while the UE remains in the RRC inactive. Rel 18 specified MT-SDT that allows the network to trigger the transmission with two mechanisms, i.e., random-access SDT and configured grant SDT. Section 5.4 shows the NR support for UAVs feature. Integrating UAVs requires careful study of UL/DL interference and mobility issues, considering their increased delay, limited MIMO capabilities, and interference to the network for ground-based UEs. Section 5.5 focuses on the study of LP-WUS/LP-WUR aiming at optimizing UE power consumption. 3GPP Rel 19 will standardize the LP-WUS design to support LP-WUR in both UE idle/inactive and connected states. Finally, Section 5.6 outlines the DSS feature, which ensures smooth transitions between LTE and NR technologies. Rel 18 adds support for handling strong interference from neighboring LTE cells and increases the downlink control channel capacity and coverage. Building on the significant advancements made in Rel 15 through Rel 18, DSS has reached a stage where further enhancements are not deemed essential for the 3GPP Rel 19. 3GPP Rel 18 also indicates the enhancements for other existing features, e.g., edge computing, network slicing, MBS, and NTN integration, and introduces new services, e.g., digital twins and extended reality.

Although still under development, Rel 18 lays the groundwork for a smarter, more adaptable, and versatile mobile network. Rel 19 and beyond will refine and expand these advancements, further solidifying the role of 5G Advanced in powering a connected future.

6. Release 19 (as Planned) and Beyond

3GPP took the Rel 19 plenary workshops in June 2023 and December 2023, which dealt with the enhancement of 5G Advanced and the preparation of 6G [210]. It has been announced that 3GPP will host the 6G workshop at the start of Rel 20 before delving into the 6G study. Moreover, Rel 21 will mark the transition from 5G Advanced to the initial stage of 6G. This section briefly overviews the heightened expectations for 5G RAN from the Rel 19 workshop, which is expected to fulfill the commercial deployment needs of 5G Advanced. The key aspects include AI/ML enhancement for NG RAN, integrated sensing and communication (ISaC), ambient IoT, and NTN evolution.

6.1. AI/ML Enhancement for NG RAN

In 3GPP Rel 19, a deeper exploration of use cases is anticipated to facilitate AI/ML applications better. For instance, applications involving fast adaptation and distributed learning might require direct sharing of data and AI/ML models between devices without traversing 5G networks [211]. Moreover, within the context of distributed learning, 5G systems must efficiently manage scenarios such as device mobility (in and out of coverage areas) [212], model transition [213, 214], energy savings [214], power allocation compensation [215], computational offloading between devices, network slicing availability [216], activation and deactivation of secondary cell groups [215], and the trade-off between AI/ML model accuracy, model generation latency, power constraints, and computing capabilities.

6.2. Integrated Sensing and Communication

3GPP Rel 19 describes use cases and potential requirements for enhancing 5G advanced to provide sensing services addressing different target verticals [217, 218]. First, use cases for ISaC can be categorized as indoor environments (home, office, and factory), highway scenarios (automotive, traffic monitoring, and intrusion detection), high-speed railway applications (autopilot and intrusion detection), weather forecasting (rainfall and flooding), UAV operations (flight trajectory tracing, UAV collision, and intrusion detection), traffic management (tourist/sports hotspot detection and car parking), health monitoring (heart rate, breathing, and sleeping), and extended reality experiences (gaming and metaverse) [219]. Subsequently, in Rel 19, KPIs are determined for each use case, relating to confidence level, the accuracy of positioning/velocity estimates through sensing (horizontal and vertical), sensing resolution (range and velocity resolution), maximum sensing service latency, refresh rate, missed detection, and false alarms [217, 219].

Notably, when delving into ISaC, the conventional 3GPP-based channel models and methodologies in [220] cannot be utilized to evaluate ISaC performance. The channel models and methodologies for ISaC require updating with verification

[221], serving as the first step for ISaC in Rel 19. This may include designing clustering models, echo path loss, interference models for sensing operations, moving target models, and monostatic and bistatic sensing.

6.3. Ambient IoT

3GPP Rel 19 anticipates improvements from two sources: (*i*) ambient power-enabled IoT using the energy harvesting (EH) concept and (*ii*) the commercial relevance of ambient IoT features. Regarding ambient power-enabled IoT, the significance of EH is discussed for future IoT networks, particularly due to energy-constrained IoT devices. The investigation includes assessing the individual benefit of EH to communications, evaluating the effectiveness of simultaneous wireless information and power transfer, and characterizing EH-based IoT devices for communication optimization problems [222]. Concerning commercial aspects, Rel 19 considers extremely low-cost IoT devices and explores use cases for different types of IoT devices, such as passive, semi-passive, and active devices [223].

6.4. NTN Evolution

3GPP Rel 19 will take into account the NTN evolution from both NR NTN and IoT NTN [224]. From the perspective of NR NTN, it is expected to include coverage enhancements (both UL and DL), mobility enhancement for NTN in connected mode, alert channels for UE terminating calls, support of RedCap terminals, support for regenerative payloads, multicast and broadcast services, discontinuous coverage, network-based positioning enhancements, and asynchronous multi-connectivity. Meanwhile, IoT NTN expects NTN mobility enhancement, enhanced HARQ disabling, support of regenerative payloads, and 5GC support for IoT NTN.

Furthermore, 3GPP Rel 19 will investigate scenarios where the backhaul connection is intermittent [225]. For example, a satellite is expected to orbit the globe regularly and receive data from a location where the direct backhaul link is unavailable. In this scenario, satellite communication is required to provide hold-and-forward capability. In another scenario, a device may use a 3GPP-based satellite communication architecture to determine its locations [226].

6.5. Summary and Discussion

3GPP Rel 19, the next wave of 5G Advanced planned to be completed by the end of 2025, will primarily focus on fulfilling commercial deployment needs while serving as a stepping stone toward the upcoming 6G. The heightened expectations in 3GPP Rel 19 can be summarized as follows. Section 6.1 indicates the expectations from the AI/ML air interface to improve intelligent NG RAN in terms of mobility, model training, network management, and data collection. Section 6.2 introduces the use cases and KPIs for each use case within the ISaC scope, where sensing functionalities simultaneously performed with communication pose challenges to updating channel models and their validation. Subsequently, in Section 6.3, the ambient power-enabled IoT and the commercial relevance of ambient IoT features are discussed. Finally, Section 6.4 provides the expected NTN evolution from both NR NTN and IoT NTN, especially for having global standards for satellite communications.

The emerging use cases and requirements will present challenges that surpass the capabilities of 5G Advanced, necessitating the advent of 6G. Notably, a 7-24 GHz spectrum is actively being prepared for 6G standardization [227]. 3GPP will validate its existing channel models through measurements associated with this spectrum. Subsequently, as necessary, modifications to channel models within these bands will be investigated in at least two aspects for applicable scenarios, including near-field propagation and spatial non-stationarity. Alongside the channel model modifications required in ISaC, this poses challenges in which the modifications should be conservatively studied to fulfill 5G Advanced and realize 6G implementation. Interestingly, "there is always a next Release" is a common saying in 3GPP [228]. This implies that certain aspects of the expectations and plans for 3GPP releases may unintentionally fail, but the opportunity to address them in the next release is appreciated.

7. Concluding Remarks

Recent successes and the rapid growth of 5G networks have attracted considerable attention from the engineering community to investigate its foundational enablers thoroughly to promote the maturity of 5G technologies. This paper aims to provide state-of-the-art knowledge and developments of 5G access technologies and beyond to support these studies, which have been officially specified in 3GPP standards from Rel 15 to Rel 17 and are expected in Rel 18 and Rel 19. Initiated from Rel 15 with 5G NR, various novel technologies have been involved in enabling potential 5G services and applications according to the focus at each release. This survey is expected to equip interested engineers and scholars with a systematic reference framework for technology development and future trends.

References

[1] Ericsson mobility report (November 2023).
URL `https://www.ericsson.com/en/reports-and-papers/mobility-report`

[2] M Series, Detailed specifications of the terrestrial radio interfaces of international mobile telecommunications-2020 (IMT-2020), Report ITU-R M.2150-1, ITU Radiocommunication Sector (Feb. 2022).

[3] 3GPP, Release 15 description; summary of Rel-15 work items, Technical Report 3GPP TR 21.915, 3rd Generation Partnership Project (Oct. 2019).

[4] 3GPP, Release 16 description; summary of Rel-16 work items, Technical Report 3GPP TR 21.916, 3rd Generation Partnership Project (Jul. 2022).

[5] 3GPP, Release 17 description; summary of Rel-17 work items, Technical Report 3GPP TR 21.917, 3rd Generation Partnership Project (Sep. 2022).

[6] M. Giordani, M. Polese, A. Roy, D. Castor, M. Zorzi, A tutorial on beam management for 3GPP NR at mmWave frequencies, IEEE Communications Surveys & Tutorials 21 (1) (2019) 173–196.

[7] S. Jun, Y. Kang, J. Kim, C. Kim, Ultra-low-latency services in 5G systems: A perspective from 3GPP standards, ETRI journal 42 (5) (2020) 721–733.

[8] T.-K. Le, U. Salim, F. Kaltenberger, An overview of physical layer design for ultra-reliable low-latency communications in 3GPP releases 15, 16, and 17, IEEE access 9 (2020) 433–444.

[9] J. Cao, M. Ma, H. Li, R. Ma, Y. Sun, P. Yu, L. Xiong, A survey on security aspects for 3GPP 5G networks, IEEE communications surveys & tutorials 22 (1) (2020) 170–195.

[10] T. Jiang, J. Zhang, P. Tang, L. Tian, Y. Zheng, J. Dou, H. Asplund, L. Raschkowski, R. D'Errico, T. Jämsä, 3GPP standardized 5G channel model for IIoT scenarios: A survey, IEEE Internet of Things Journal 8 (11) (2021) 8799–8815.

[11] X. Lin, S. Rommer, S. Euler, E. A. Yavuz, R. S. Karlsson, 5g from space: An overview of 3gpp non-terrestrial networks, IEEE Communications Standards Magazine 5 (4) (2021) 147–153. doi:10.1109/mcomstd.011.2100038.

[12] Z. Ali, S. Lagén, L. Giupponi, R. Rouil, 3GPP NR V2X mode 2: Overview, models and system-level evaluation, IEEE Access 9 (2021) 89554–89579.

[13] 3GPP, Study on new radio access technology: Radio access architecture and interfaces (Release 14), Technical Report 3GPP TR 38.801, 3rd Generation Partnership Project (Mar. 2017).

[14] L. M. P. Larsen, A. Checko, H. L. Christiansen, A survey of the functional splits proposed for 5g mobile crosshaul networks, IEEE Communications Surveys & Tutorials 21 (1) (2019) 146–172. doi:10.1109/comst.2018.2868805.

[15] 3GPP, Study on CU-DU lower layer split for NR (Release 15), Technical Report 3GPP TR 38.816, 3rd Generation Partnership Project (Dec. 2017).

[16] 3GPP, NG-RAN; architecture description (Release 15), Technical Specification 3GPP TS 38.401, 3rd Generation Partnership Project (Sep. 2020).

[17] 3GPP, Study of separation of NR Control Plane (CP) and User Plane (UP) for split option 2, Technical Report 3GPP TR 38.806, 3rd Generation Partnership Project (Jan. 2018).

[18] H. He, L. Yang, L. Zhuang, H. Jiren, G. Yin, Mechanism of fast data retransmission in CU-DU split architecture of 5G NR, ZTE Communications 16 (3) (2020) 40–44. doi:10.19729/j.cnki.1673-5188.2018.03.007.

[19] S. Xu, M. Hou, Y. Fu, H. Bian, C. Gao, Improved fast centralized retransmission scheme for high-layer functional split in 5g network, Journal of Physics: Conference Series 960 (2018) 012006. doi:10.1088/1742-6596/960/1/012006.

[20] F. D. L. Coutinho, H. S. Silva, A. S. R. Oliveira, Fpga-based design and optimization of a 5g-nr du receiver, in: 2021 Telecoms Conference (ConfTELE), IEEE, 2021. doi:10.1109/conftele50222.2021.9435579.

[21] S. Kar, P. Mishra, K.-C. Wang, A novel single grant-based uplink scheme for high throughput and reliable low latency communication, in: 2023 19th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), IEEE, 2023, pp. 169–174. doi:10.1109/wimob58348.2023.10187774.

[22] M. Polese, L. Bonati, S. D'Oro, S. Basagni, T. Melodia, Understanding o-ran: Architecture, interfaces, algorithms, security, and research challenges, IEEE Communications Surveys & Tutorials 25 (2) (2023) 1376–1411. doi:10.1109/comst.2023.3239220.

[23] 3GPP, Study on self evaluation towards IMT-2020 submission (Release 15), Technical Report 3GPP TR 37.910, 3rd Generation Partnership Project (Sep. 2018).

[24] 3GPP, Physical layer procedures for data (Release 17), Technical Specification 3GPP TS 38.214, 3rd Generation Partnership Project (Dec. 2023).

[25] 3GPP, New frequency range for NR (3.3–4.2 GHz) (Release 15), Technical Report 3GPP TR 38.813, 3rd Generation Partnership Project (Mar. 2018).

[26] 3GPP, New frequency range for NR (4.4–5.0 GHz) (Release 15), Technical Report 3GPP TR 38.814, 3rd Generation Partnership Project (Jun. 2018).

[27] 3GPP, New frequency range for NR (24.25–29.5 GHz) (Release 15), Technical Report 3GPP TR 38.815, 3rd Generation Partnership Project (Sep. 2021).

[28] Federal Communication Commission, Report and order FCC-16-89 (Jul. 2016).
URL https://apps.fcc.gov/edocs_public/attachmatch/FCC-16-89A1.pdf

[29] 3GPP, 7 - 24 GHz frequency range (Release 16), Technical Report 3GPP TR 38.820, 3rd Generation Partnership Project (Mar. 2021).

[30] 3GPP, Study on supporting NR from 52.6 GHz to 71 GHz (Release 17), Technical Report 3GPP TR 38.808, 3rd Generation Partnership Project (Mar. 2021).

[31] Qualcomm, Engaging 60 GHz and beyond in 5G (Jan. 2020).
URL https://rww2020.iot.ieee.org/wp-content/uploads/sites/124/2020/01/01-26-2020-Xiaoxia-Zhang-Qualcomm-Presentation.pdf

[32] B. Coll-Perales, J. Gozalvez, M. Gruteser, Sub-6ghz assisted mac for millimeter wave vehicular communications, IEEE Communications Magazine 57 (3) (2019) 125–131. doi:10.1109/mcom.2019.1800509.

[33] M. Zada, I. A. Shah, H. Yoo, Integration of sub-6-ghz and mm-wave bands with a large frequency ratio for future 5g mimo applications, IEEE Access 9 (2021) 11241–11251. doi:10.1109/access.2021.3051066.

[34] L. Sang, Z. Hu, S. Wu, W. Huang, H. Tu, W. Wang, P. Chen, W. Hong, A dual-band planar antenna array with high-frequency ratio for both sub-6 band and mm-waveband, IEEE Transactions on Antennas and Propagation 71 (5) (2023) 3856–3867. doi:10.1109/tap.2023.3247170.

[35] S. Islam, M. Zada, H. Yoo, Low-pass filter based integrated 5g smartphone antenna for sub-6-ghz and mm-wave bands, IEEE Transactions on Antennas and Propagation 69 (9) (2021) 5424–5436. doi:10.1109/tap.2021.3061012.

[36] H. Miao, J. Zhang, P. Tang, L. Tian, X. Zhao, B. Guo, G. Liu, Sub-6 ghz to mmwave for 5g-advanced and beyond: Channel measurements, characteristics and impact on system performance, IEEE Journal on Selected Areas in Communications 41 (6) (2023) 1945–1960. doi:10.1109/jsac.2023.3274175.

[37] 3GPP, Physical channels and modulation (Release 15), Technical Specification 3GPP TS 38.211, 3rd Generation Partnership Project (Dec. 2021).

[38] 3GPP, NR: Physical layer procedures for data (Release 15), Technical Specification 3GPP TS 38.214, 3rd Generation Partnership Project (Mar. 2022).

[39] 3GPP, Physical channels and modulation (Release 16), Technical Specification 3GPP TS 38.211, 3rd Generation Partnership Project (Sep. 2018).

[40] S. Kutty, D. Sen, Beamforming for millimeter wave communications: An inclusive survey, IEEE Communications Surveys & Tutorials 18 (2) (2016) 949–973. doi:10.1109/comst.2015.2504600.

[41] M. Giordani, M. Polese, A. Roy, D. Castor, M. Zorzi, A tutorial on beam management for 3gpp nr at mmwave frequencies, IEEE Communications Surveys & Tutorials 21 (1) (2019) 173–196. doi:10.1109/comst.2018.2869411.

[42] 3GPP, NR: Physical layer procedures for data (Release 16), Technical Specification 3GPP TS 38.214, 3rd Generation Partnership Project (Apr. 2021).

[43] 3GPP, Study on New Radio (NR) access technology, Technical Report 3GPP TR 38.912, 3rd Generation Partnership Project (Sep. 2018).

[44] I. Ahmed, H. Khammari, A. Shahid, A. Musa, K. S. Kim, E. De Poorter, I. Moerman, A survey on hybrid beamforming techniques in 5G: Architecture and system model perspectives, IEEE Communications Surveys & Tutorials 20 (4) (2018) 3060–3097.

[45] F. W. Vook, W. J. Hillery, E. Visotsky, J. Tan, X. Shao, M. Enescu, System level performance characteristics of sub-6ghz massive mimo deployments with the 3gpp new radio, in: 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall), IEEE, 2018. doi:10.1109/vtcfall.2018.8690560.

[46] L. Ge, Y. Zhang, G. Chen, J. Tong, Compression-based lmmse channel estimation with adaptive sparsity for massive mimo in 5g systems, IEEE Systems Journal 13 (4) (2019) 3847–3857. doi:10.1109/jsyst.2019.2897862.

[47] Z. Liu, S. Sun, Q. Gao, H. Li, CSI feedback based on spatial and frequency domains compression for 5G multi-user massive MIMO systems, in: 2019 IEEE/CIC International Conference on Communications in China (ICCC), IEEE, 2019, pp. 834–839. doi:10.1109/iccchina.2019.8855979.

[48] D. F. Carrera, D. Zabala-Blanco, C. Vargas-Rosales, C. A. Azurdia-Meza, Extreme learning machine-based receiver for multi-user massive mimo systems, IEEE Communications Letters 25 (2) (2021) 484–488.

doi:10.1109/lcomm.2020.3031195.

[49] 3GPP, Study on scenarios and requirements for next generation access technologies (Release 15), Technical Report 3GPP TR 38.913, 3rd Generation Partnership Project (Jun. 2018).

[50] 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, M. Windisch, Latency critical iot applications in 5g: Perspective on the design of radio interface and network architecture, IEEE Communications Magazine 55 (2) (2017) 70–78. doi:10.1109/mcom.2017.1600435cm.

[51] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, H. Dai, A survey on low latency towards 5g: Ran, core network and caching solutions, IEEE Communications Surveys & Tutorials 20 (4) (2018) 3098–3130. doi:10.1109/comst.2018.2841349.

[52] ITU-T Technology Watch Report, The tactile Internet (Aug. 2014). URL https://www.itu.int/oth/T2301000023/en

[53] 3GPP, Study on enhancement of 3GPP support for 5G V2X services (Release 15), Technical Report 3GPP TR 22.886, 3rd Generation Partnership Project (Sep. 2018).

[54] D. Soldani, Y. J. Guo, B. Barani, P. Mogensen, C.-L. I, S. K. Das, 5g for ultra-reliable low-latency communications, IEEE Network 32 (2) (2018) 6–7. doi:10.1109/mnet.2018.8329617.

[55] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, B. Shim, Ultra-reliable and low-latency communications in 5g downlink: Physical layer aspects, IEEE Wireless Communications 25 (3) (2018) 124–130. doi:10.1109/mwc.2018.1700294.

[56] N. H. Tu, K. Lee, Performance analysis and optimization of multihop mimo relay networks in short-packet communications, IEEE Transactions on Wireless Communications 21 (6) (2022) 4549–4562. doi:10.1109/twc.2021.3131205.

[57] P. Popovski, C. Stefanovic, J. J. Nielsen, E. de Carvalho, M. Angjelichinoski, K. F. Trillingsgaard, A.-S. Bana, Wireless access in ultra-reliable low-latency communication (urllc), IEEE Transactions on Communications 67 (8) (2019) 5783–5801. doi:10.1109/tcomm.2019.2914652.

[58] J. Sachs, G. Wikstrom, T. Dudda, R. Baldemair, K. Kittichokechai, 5g radio network design for ultra-reliable low-latency communication, IEEE Network 32 (2) (2018) 24–31. doi:10.1109/mnet.2018.1700232.

[59] A. A. Esswie, K. I. Pedersen, On the ultra-reliable and low-latency communications in flexible tdd/fdd 5g networks, in: 2020 IEEE 17th Annual Consumer Communications & Networking Conference (CCNC), IEEE, 2020. doi:10.1109/ccnc46108.2020.9045657.

[60] 3GPP, Study on enhancement of 3GPP support for 5G V2X services (Release 16), Technical Report 3GPP TR 22.886, 3rd Generation Partnership Project (Dec. 2018).

[61] Y. Chen, A. Bayesteh, Y. Wu, B. Ren, S. Kang, S. Sun, Q. Xiong, C. Qian, B. Yu, Z. Ding, S. Wang, S. Han, X. Hou, H. Lin, R. Visoz, R. Razavi, Toward the standardization of non-orthogonal multiple access for next generation wireless networks, IEEE Communications Magazine 56 (3) (2018) 19–27. doi:10.1109/mcom.2018.1700845.

[62] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, I. Chih-Lin, H. V. Poor, Application of non-orthogonal multiple access in lte and 5g networks, IEEE Communications Magazine 55 (2) (2017) 185–191. doi:10.1109/mcom.2017.1500657cm.

[63] K. Yang, N. Yang, N. Ye, M. Jia, Z. Gao, R. Fan, Non-orthogonal multiple access: Achieving sustainable future radio access, IEEE Communications Magazine 57 (2) (2019) 116–121. doi:10.1109/mcom.2018.1800179.

[64] Y. Yuan, Z. Yuan, G. Yu, C.-h. Hwang, P.-k. Liao, A. Li, K. Takeda, Non-orthogonal transmission technology in lte evolution, IEEE Communications Magazine 54 (7) (2016) 68–74. doi:10.1109/mcom.2016.7509381.

[65] 3GPP, Study on non-orthogonal multiple access (NOMA) for NR (Release 16), Technical Specification 3GPP TR 38.812, 3rd Generation Partnership Project (Jul. 2020).

[66] L. Zhang, X. Xu, Y. Chen, W. Yiqun, Y. Du, Grant-free transmission method, terminal, and network device, uS Patent App. 16/451,728 (Oct. 2019).

[67] NTT DoCoMo, New uplink non-orthogonal multiple access schemes for NR, Discussion, Decision 3GPP R1-165174, 3GPP TSG RAN WG1 #85, Nanjing, China (May 2016).

[68] NTT DoCoMo, Initial views and evaluation results on non-orthogonal multiple access for NR, Discussion, Decision 3GPP R1-165175, 3GPP TSG RAN WG1 #85, Nanjing, China (May 2016).

[69] Nokia, Performance of interleave division multiple access (IDMA) in combination with OFDM family waveforms, Discussion, Decision 3GPP R1-165021, 3GPP TSG RAN WG1 #85, Nanjing, China (May 2016).

[70] Samsung, Non-orthogonal multiple access candidate for NR, Discussion, Decision 3GPP R1-163992, 3GPP TSG RAN WG1 #85, Nanjing, China (May 2016).

[71] Samsung, Low code rate and signature based multiple access scheme for new radio, Discussion, Decision 3GPP R1-164869, 3GPP TSG RAN WG1 #85, Nanjing, China (May 2016).

[72] Intel, Multiple access schemes for new radio interface, Discussion, Decision 3GPP R1-162385, 3GPP TSG RAN WG1 #84bis, Busan, South Korea (Apr. 2016).

[73] Huawei, HiSilicon, Overview of non-orthogonal multiple access for 5G, Discussion, Decision 3GPP R1-162153, 3GPP TSG RAN WG1 #84bis, Busan, Korea (Apr. 2016).

[74] CATT, Candidate solution for new multiple access, Discussion, Decision 3GPP R1-163383, 3GPP TSG RAN WG1 #84bis, Busan, Korea (Apr. 2016).

[75] Fujitsu, Initial lls results for UL non-orthogonal multiple acces, Discussion, Decision 3GPP R1-164329, 3GPP TSG RAN WG1 #85, Nanjing, China (May 2016).

[76] ZTE, Grant-free multiple access for mMTC based on short spreading, Discussion, Decision 3GPP R1-164269, 3GPP TSG RAN WG1 #85, Nanjing, China (May 2016).

[77] LG Electronics, Considerations on DL/UL multiple access for NR, Discussion, Decision 3GPP R1-162517, 3GPP TSG RAN WG1 #84bis, Busan, Korea (Apr. 2016).

[78] Nokia, Alcatel-Lucent Shanghai Bell, Non-orthogonal multiple access for new radio, Discussion, Decision 3GPP R1-165019, 3GPP TSG-RAN WG1 #85, Nanjing, China (May 2016).

[79] Qualcomm, Resource spread multiple access, Discussion, Decision 3GPP R1-164688, 3GPP TSG RAN WG1 #85, Nanjing, China (May 2016).

[80] MediaTek, New uplink non-orthogonal multiple access schemes for NR, Discussion, Decision 3GPP R1-167535, 3GPP TSG RAN WG1 #86, Gothenburg, Sweden (Aug. 2016).

[81] K. HIGUCHI, A. BENJEBBOUR, Non-orthogonal multiple access (noma) with successive interference cancellation for future radio access, IEICE Transactions on Communications E98.B (3) (2015) 403–414. doi:10.1587/transcom.e98.b.403.

[82] NTT DoCoMo, Candidate solution for new multiple access, Discussion, Decision 3GPP R1-163111, 3GPP TSG RAN WG1 #84bis, Busan, Korea (Apr. 2016).

[83] Z. Ding, Z. Yang, P. Fan, H. V. Poor, On the performance of non-orthogonal multiple access in 5g systems with randomly deployed users, IEEE Signal Processing Letters 21 (12) (2014) 1501–1505. doi:10.1109/lsp.2014.2343971.

[84] Y. Zhang, H.-M. Wang, T.-X. Zheng, Q. Yang, Energy-efficient transmission design in non-orthogonal multiple access, IEEE Transactions on Vehicular Technology 66 (3) (2017) 2852–2857. doi:10.1109/tvt.2016.2578949.

[85] M. Zeng, A. Yadav, O. A. Dobre, H. V. Poor, Energy-efficient power allocation for uplink noma, in: 2018 IEEE Global Communications Conference (GLOBECOM), IEEE, 2018. doi:10.1109/glocom.2018.8647478.

[86] A. Haghighat, S. N. Nazar, S. Herath, R. Olesen, On the performance of idma-based non-orthogonal multiple access schemes, in: 2017 IEEE 86th Vehicular Technology Conference (VTC-Fall), IEEE, 2017. doi:10.1109/vtcfall.2017.8288410.

[87] M. Vaezi, Z. Ding, H. V. Poor, Multiple Access Techniques for 5G Wireless Networks and Beyond, Vol. 159, Springer International Publishing, 2019. doi:10.1007/978-3-319-92090-0.

[88] S. Hu, B. Yu, C. Qian, Y. Xiao, Q. Xiong, C. Sun, Y. Gao, Nonorthogonal interleave-grid multiple access scheme for industrial internet of things in 5g network, IEEE Transactions on Industrial Informatics 14 (12) (2018) 5436–5446. doi:10.1109/tii.2018.2858142.

[89] Samsung, Link level performance evaluation for IGMA, Discussion, De-

cision 3GPP R1-166750, 3GPP TSG RAN WG1 #86, Gothenburg, Sweden (Aug. 2016).

[90] H. Nikopour, H. Baligh, Sparse code multiple access, in: 2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), IEEE, 2013, pp. 332–336. doi:10.1109/pimrc.2013.6666156.

[91] H. Yu, Z. Fei, N. Yang, N. Ye, Optimal design of resource element mapping for sparse spreading non-orthogonal multiple access, IEEE Wireless Communications Letters 7 (5) (2018) 744–747. doi:10.1109/lwc.2018.2818157.

[92] S. Chaturvedi, Z. Liu, V. A. Bohara, A. Srivastava, P. Xiao, A tutorial on decoding techniques of sparse code multiple access, IEEE Access 10 (2022) 58503–58524. doi:10.1109/access.2022.3178127.

[93] P. Li, Y. Jiang, S. Kang, F. Zheng, X. You, Pattern division multiple access with large-scale antenna array, in: 2017 IEEE 85th Vehicular Technology Conference (VTC Spring), IEEE, 2017. doi:10.1109/vtcspring.2017.8108482.

[94] J. Zeng, B. Liu, X. Su, Interleaver-based pattern division multiple access with iterative decoding and detection, in: 2017 IEEE 85th Vehicular Technology Conference (VTC Spring), IEEE, 2017. doi:10.1109/vtcspring.2017.8108432.

[95] X. Dai, Z. Zhang, B. Bai, S. Chen, S. Sun, Pattern division multiple access: A new multiple access technology for 5g, IEEE Wireless Communications 25 (2) (2018) 54–60. doi:10.1109/mwc.2018.1700084.

[96] J. Zhang, X. Wang, X. Yang, H. Zhou, Low density spreading signature vector extension (lds-sve) for uplink multiple access, in: 2017 IEEE 86th Vehicular Technology Conference (VTC-Fall), IEEE, 2017. doi:10.1109/vtcfall.2017.8287908.

[97] Z. Yuan, G. Yu, W. Li, Y. Yuan, X. Wang, J. Xu, Multi-user shared access for internet of things, in: 2016 IEEE 83rd Vehicular Technology Conference (VTC Spring), IEEE, 2016. doi:10.1109/vtcspring.2016.7504361.

[98] A. Medra, T. N. Davidson, Flexible codebook design for limited feedback systems via sequential smooth optimization on the grassmannian manifold, IEEE Transactions on Signal Processing 62 (5) (2014) 1305–1318. doi:10.1109/tsp.2014.2301137.

[99] Qualcomm, Candidate NR multiple access schemes, Discussion, Decision 3GPP R1-162202, 3GPP TSG RAN WG1 #84bis, Busan, South Korea (Apr. 2016).

[100] P. Li, J. Xu, Uav-enabled cellular networks with multi-hop backhauls: Placement optimization and wireless resource allocation, in: 2018 IEEE International Conference on Communication Systems (ICCS), IEEE, 2018, pp. 110–114. doi:10.1109/iccs.2018.8689218.

[101] B. Galkin, J. Kibilda, L. A. DaSilva, Backhaul for low-altitude uavs in urban environments, in: 2018 IEEE International Conference on Communications (ICC), IEEE, 2018. doi:10.1109/icc.2018.8422376.

[102] 3GPP, LTE general packet radio service (GPRS) enhancements for evolved universal terrestrial radio access network (E-UTRAN) access (Release 10), Technical Report 3GPP TS 23.401, 3rd Generation Partnership Project (Mar. 2011).

[103] M. N. Islam, S. Subramanian, A. Sampath, Integrated access backhaul in millimeter wave networks, in: 2017 IEEE Wireless Communications and Networking Conference (WCNC), IEEE, 2017. doi:10.1109/wcnc.2017.7925837.

[104] 3GPP, NG-RAN architecture description (Release 16), Technical Report 3GPP TR 38.401, 3rd Generation Partnership Project (Nov. 2020).

[105] 3GPP, NG-RAN study on new radio access technology: Radio access architecture and interfaces (Release 14), Technical Report 3GPP TR 38.801, 3rd Generation Partnership Project (Mar. 2017).

[106] 3GPP, NG-RAN study on integrated access and backhaul (Release 16), Technical Report 3GPP TR 38.874, 3rd Generation Partnership Project (Dec. 2017).

[107] J. A. del Peral-Rosado, F. Gunnarsson, S. Dwivedi, S. M. Razavi, O. Renaudin, J. A. Lopez-Salcedo, G. Seco-Granados, Exploitation of 3d city maps for hybrid 5g rtt and gnss positioning simulations, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 9205–9209. doi:10.1109/icassp40776.2020.9053157.

[108] 3GPP, NG-RAN stage 2 functional specification of user equipment (UE) positioning in NG-RAN (Release 16), Technical Report 3GPP TR 38.305, 3rd Generation Partnership Project (Jul. 2020).

[109] 3GPP, NG-RAN NR positioning protocol a (NRPPa) (Release 16), Technical Report 3GPP TR 38.305, 3rd Generation Partnership Project (Nov. 2020).

[110] 3GPP, 5G:study on NR positioning support (Release 16), Technical Specification 3GPP TS 38.855, 3rd Generation Partnership Project (Mar. 2019).

[111] F. Gustafsson, F. Gunnarsson, Positioning using time-difference of arrival measurements, in: 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)., ICASSP-03, IEEE, 2003, pp. VI–553–VI–556. doi:10.1109/icassp.2003.1201741.

[112] 3GPP, LTE location measurement unit (LMU) conformance specification; network based positioning systems in evolved universal terrestrial radio access network (E-UTRAN) (Release 16), Technical Report 3GPP TR 36.112, 3rd Generation Partnership Project (Jul. 2020).

[113] W. M. Gifford, D. Dardari, M. Z. Win, The impact of multipath information on time-of-arrival estimation, IEEE Transactions on Signal Processing 70 (2022) 31–46. doi:10.1109/tsp.2020.3038254.

[114] A. Bergstrom, G. Hendeby, F. Gunnarsson, F. Gustafsson, Toa estimation improvements in multipath environments by measurement error models, in: 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), IEEE, 2017. doi:10.1109/pimrc.2017.8292377.

[115] J. Zhou, L. Shen, Z. Sun, A new method of d-tdoa time measurement based on rtt, MATEC Web of Conferences 207 (2018) 03018. doi:10.1051/matecconf/201820703018.

[116] K. Radnosrati, C. Fritsche, F. Gunnarsson, F. Gustafsson, G. Hendeby, Localization in 3gpp lte based on one rtt and one tdoa observation, IEEE Transactions on Vehicular Technology 69 (3) (2020) 3399–3411. doi:10.1109/tvt.2020.2968118.

[117] S. He, S.-H. G. Chan, Wi-fi fingerprint-based indoor positioning: Recent advances and comparisons, IEEE Communications Surveys & Tutorials 18 (1) (2016) 466–490. doi:10.1109/comst.2015.2464084.

[118] R. Schmidt, Multiple emitter location and signal parameter estimation, IEEE Transactions on Antennas and Propagation 34 (3) (1986) 276–280. doi:10.1109/tap.1986.1143830.

[119] T. Kailath, Esprit–estimation of signal parameters via rotational invariance techniques, Optical Engineering 29 (4) (1990) 296. doi:10.1117/12.55606.

[120] X. Wu, W.-P. Zhu, J. Yan, Direction of arrival estimation for off-grid signals based on sparse bayesian learning, IEEE Sensors Journal 16 (7) (2016) 2004–2016. doi:10.1109/jsen.2015.2508059.

[121] Z.-M. Liu, Z.-T. Huang, Y.-Y. Zhou, Array signal processing via sparsity-inducing representation of the array covariance matrix, IEEE Transactions on Aerospace and Electronic Systems 49 (3) (2013) 1710–1724. doi:10.1109/taes.2013.6558014.

[122] P. Stoica, A. Nehorai, Music, maximum likelihood, and cramer-rao bound, IEEE Transactions on Acoustics, Speech, and Signal Processing 37 (5) (1989) 720–741. doi:10.1109/29.17564.

[123] D. T. Hoang, K. Lee, Deep learning-aided coherent direction-of-arrival estimation with the ftmr algorithm, IEEE Transactions on Signal Processing 70 (2022) 1118–1130. doi:10.1109/tsp.2022.3144033.

[124] D. T. Hoang, K. Lee, Coherent signal enumeration based on deep learning and the ftmr algorithm, in: ICC 2022 - IEEE International Conference on Communications, IEEE, 2022, pp. 5098–5103. doi:10.1109/icc45855.2022.9838253.

[125] W. Du, R. Kirlin, Improved spatial smoothing techniques for doa estimation of coherent signals, IEEE Transactions on Signal Processing 39 (5) (1991) 1208–1210. doi:10.1109/78.80975.

[126] W. Zhang, Y. Han, M. Jin, X.-S. Li, An improved esprit-like algorithm for coherent signals doa estimation, IEEE Communications Letters 24 (2) (2020) 339–343. doi:10.1109/lcomm.2019.2953851.

[127] W. Zhang, Y. Han, M. Jin, X. Qiao, Multiple-toeplitz matrices reconstruction algorithm for doa estimation of coherent signals, IEEE Access 7 (2019) 49504–49512. doi:10.1109/access.2019.2909783.

[128] D. T. Hoang, K. Lee, Deep learning-aided signal enumeration for lens antenna array, IEEE Access 10 (2022) 123835–123846. doi:10.1109/access.2022.3224608.

[129] T.-D. Hoang, X. Huang, P. Qin, Gradient descent-based direction-of-arrival estimation for lens antenna array, IEEE Signal Processing Letters 30 (2023) 838–842. doi:10.1109/lsp.2023.3292742.

[130] Z. Papp, G. Irvine, R. Smith, F. Mogyorosi, P. Revisnyei, I. Toros, A. Pasic, Tdoa based indoor positioning over small cell 5g networks, in: NOMS 2022-2022 IEEE/IFIP Network Operations and Management Symposium, IEEE, 2022. doi:10.1109/noms54207.2022.9789712.

[131] E. Y. Menta, N. Malm, R. Jantti, K. Ruttik, M. Costa, K. Leppanen, On the performance of aoa–based localization in 5g ultra–dense networks, IEEE Access 7 (2019) 33870–33880. doi:10.1109/access.2019.2903633.

[132] A. Kakkavas, M. H. Castaneda Garcia, R. A. Stirling-Gallacher, J. A. Nossek, Multi-array 5g v2v relative positioning: Performance bounds, in: 2018 IEEE Global Communications Conference (GLOBECOM), IEEE, 2018. doi:10.1109/glocom.2018.8647812.

[133] M. Malmstrom, I. Skog, S. M. Razavi, Y. Zhao, F. Gunnarsson, 5g positioning - a machine learning approach, in: 2019 16th Workshop on Positioning, Navigation and Communications (WPNC), IEEE, 2019. doi:10.1109/wpnc47567.2019.8970186.

[134] M. Z. Comiter, M. B. Crouse, H. T. Kung, A data-driven approach to localization for high frequency wireless mobile networks, in: GLOBECOM 2017 - 2017 IEEE Global Communications Conference, IEEE, 2017. doi:10.1109/glocom.2017.8254732.

[135] 3GPP, Study on NR-based access to unlicensed spectrum (Release 16), Technical Specification 3GPP TR 38.889, 3rd Generation Partnership Project (Nov. 2018).

[136] M. S. Ali, E. Hossain, D. I. Kim, Lte/lte-a random access for massive machine-type communications in smart cities, IEEE Communications Magazine 55 (1) (2017) 76–83. doi:10.1109/mcom.2017.1600215cm.

[137] E. Peralta, T. Levanen, F. Frederiksen, M. Valkama, Two-step random access in 5g new radio: Channel structure design and performance, in: 2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring), IEEE, 2021. doi:10.1109/vtc2021-spring51267.2021.9449057.

[138] 3GPP, Study on user equipment (UE) power saving in NR, Technical Report 3GPP TR 38.840, 3rd Generation Partnership Project (Jun. 2019).

[139] C. S. Bontu, E. Illidge, Drx mechanism for power saving in lte, IEEE Communications Magazine 47 (6) (2009) 48–55. doi:10.1109/mcom.2009.5116800.

[140] X. Lin, D. Yu, H. Wiemann, A Primer on Bandwidth Parts in 5G New Radio, Springer International Publishing, 2021, Ch. 12, pp. 357–370. doi:10.1007/978-3-030-58197-8_12.

[141] S. He, S.-H. G. Chan, Wi-fi fingerprint-based indoor positioning: Recent advances and comparisons, IEEE Communications Surveys & Tutorials 18 (1) (2016) 466–490. doi:10.1109/comst.2015.2464084.

[142] 3GPP, Multiplexing and channel coding (Release 16), Technical Report 3GPP TR 38.840, 3rd Generation Partnership Project (Dec. 2019).

[143] 3GPP, LTE: Evolved universal terrestrial radio access (E-UTRA) (Release 15), Technical Specification 3GPP TS 36.321, 3rd Generation Partnership Project (Aug. 2016).

[144] 3GPP, 5G: Physical layer procedures for control (Release 16), Technical Specification 3GPP TS 36.213, 3rd Generation Partnership Project (Jul. 2020).

[145] Ang, Peter Pui Lok and Sarkis, Gabi and Gaal, Peter and Chen, Wanshi and Soriaga, Joseph Binamira and Lee, Heechoon and Hosseini, Seyedkianoush and Xu, Huilin and Nam, Wooseok and Ly, Hung Dinh, Secondary cell dormancy for new radio carrier aggregation, U.S. Patent App. 16/739,035 (Jul. 2020).

[146] 3GPP, On adaptation aspects for NR UE power consumption reduction, Technical Report R1-1812421, ZTE (Nov. 2018).

[147] 3GPP, Service Requirements for Enhanced V2X Scenarios (Release 16), Technical Report 3GPP TR 22.186, 3rd Generation Partnership Project (Jun. 2019).

[148] 3GPP, Requirements for support of radio resource management (Release 16), Technical Specification 3GPP TS 38.133, 3rd Generation Partnership Project (Sep. 2020).

[149] 3GPP, Summary of RAN1 Agreements/Working Assumptions in WI 5G V2X With NR Sidelink, Technical Report R1-1913601, LG Electron. (Nov. 2019).

[150] 3GPP, Overall description of radio access network (RAN) aspects for vehicle-to-everything (V2X) based on LTE and NR (Release 16), Technical Report 3GPP TR 37.985, 3rd Generation Partnership Project (Sep. 2020).

[151] 3GPP, Study on NR-based access to unlicensed spectrum (Release 16), Technical Report 3GPP TR 38.889, 3rd Generation Partnership Project (Dec. 2018).

[152] Qualcomm, Summary for WI on NR-based access to unlicensed spectrum, Discussion, Decision RP-202753, 3GPP TSG-RAN Meeting #90e, E-Meeting (Dec. 2020).

[153] 3GPP TSG RAN, New WID on NR-Based Access to Unlicensed Spectrum, Approval RP-182878, 3GPP TSG-RAN Meeting #82, Meeting (Dec. 2018).

[154] S. Muhammad, H. H. Refai, M. O. Al Kalaa, 5g nr-u: Homogeneous coexistence analysis, in: GLOBECOM 2020 - 2020 IEEE Global Communications Conference, IEEE, 2020. doi:10.1109/globecom42002.2020.9322216.

[155] G. Naik, J. Liu, J.-M. Park, Coexistence of wireless technologies in the 5 ghz bands: A survey of existing solutions and a roadmap for future research, IEEE Communications Surveys & Tutorials 20 (3) (2018) 1777–1798. doi:10.1109/comst.2018.2815585.

[156] K. Kosek-Szott, A. Lo Valvo, S. Szott, P. Gallo, I. Tinnirello, Downlink channel access performance of nr-u: Impact of numerology and mini-slots on coexistence with wi-fi in the 5 ghz band, Computer Networks 195 (2021) 108188. doi:10.1016/j.comnet.2021.108188.

[157] L. Wang, M. Zeng, J. Guo, Q. Cui, Z. Fei, Joint bandwidth and transmission opportunity allocation for the coexistence between nr-u and wifi systems in the unlicensed band, IEEE Transactions on Vehicular Technology 70 (11) (2021) 11881–11893. doi:10.1109/tvt.2021.3116378.

[158] Q. Ren, J. Zheng, B. Wang, Y. Zhang, Performance modeling of an nr-u and wifi coexistence system with nr-u type b multichannel access procedure, IEEE Internet of Things Journal 10 (5) (2023) 4403–4419. doi:10.1109/jiot.2022.3218068.

[159] V. Loginov, A. Troegubov, A. Lyakhov, E. Khorov, Enhanced collision resolution methods with mini-slot support for 5g nr-u, IEEE Access 9 (2021) 146137–146152. doi:10.1109/access.2021.3122953.

[160] V. Loginov, E. Khorov, A. Lyakhov, I. F. Akyildiz, Cr-lbt: Listen-before-talk with collision resolution for 5g nr-u networks, IEEE Transactions on Mobile Computing 21 (9) (2022) 3138–3149. doi:10.1109/tmc.2021.3055028.

[161] T.-K. Le, F. Kaltenberger, U. Salim, Dynamic switch between load based and frame based channel access mechanisms in unlicensed spectrum, in: GLOBECOM 2021 - 2021 IEEE Global Communications Conference, IEEE, 2021. doi:10.1109/GLOBECOM46510.2021.9685120.

[162] M. Lauridsen, D. Laselva, F. Frederiksen, J. Kaikkonen, 5g new radio user equipment power modeling and potential energy savings, in: 2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall), IEEE, 2019. doi:10.1109/vtcfall.2019.8891215.

[163] F. Wang, D. Guan, L. Zhao, K. Zheng, Cooperative v2x for high definition map transmission based on vehicle mobility, in: 2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring), IEEE, 2019. doi:10.1109/vtcspring.2019.8746537.

[164] H. Khan, S. Samarakoon, M. Bennis, Enhancing video streaming in vehicular networks via resource slicing, IEEE Transactions on Vehicular Technology 69 (4) (2020) 3513–3522. doi:10.1109/tvt.2020.2975068.

[165] 3GPP, Study on support of reduced capability NR devices, Technical Report 3GPP TR 38.875, 3rd Generation Partnership Project (Mar. 2021).

[166] S. N. K. Veedu, M. Mozaffari, A. Hoglund, E. A. Yavuz, T. Tirronen, J. Bergman, Y.-P. E. Wang, Toward smaller and lower-cost 5g devices with longer battery life: An overview of 3gpp release 17 redcap, IEEE Communications Standards Magazine 6 (3) (2022) 84–90. doi:10.1109/mcomstd.0001.2200029.

[167] N.-N. Dao, Internet of wearable things: Advancements and benefits from 6g technologies, Future Generation Computer Systems 138 (2023) 172–184. doi:10.1016/j.future.2022.07.006.

[168] S. Moloudi, M. Mozaffari, S. N. K. Veedu, K. Kittichokechai, Y.-P. E. Wang, J. Bergman, A. Hoglund, Coverage evaluation for 5g reduced capability new radio (nr-redcap), IEEE Access 9 (2021) 45055–45067. doi:10.1109/access.2021.3066036.

[169] 3GPP, Study on narrow-band internet of things (NB-IoT) / enhanced machine type communication (eMTC) support for non-terrestrial networks (NTN), Technical Report 3GPP TR 36.763, 3rd Generation Partnership Project (Jun. 2021).

[170] X. Lin, Z. Lin, S. E. Lowenmark, J. Rune, R. Karlsson, Ericsson, Doppler shift estimation in 5g new radio non-terrestrial networks, in: 2021 IEEE Global Communications Conference (GLOBECOM), IEEE, 2021. doi:10.1109/globecom46510.2021.9685184.

[171] N.-N. Dao, Q.-V. Pham, N. H. Tu, T. T. Thanh, V. N. Q. Bao, D. S. Lakew, S. Cho, Survey on aerial radio access networks: Toward a comprehensive 6g access infrastructure, IEEE Communications Surveys & Tutorials 23 (2) (2021) 1193–1225. doi:10.1109/comst.2021.3059644.

[172] M. M. Azari, S. Solanki, S. Chatzinotas, O. Kodheli, H. Sallouha, A. Colpaert, J. F. Mendoza Montoya, S. Pollin, A. Haqiqatnejad, A. Mostaani, E. Lagunas, B. Ottersten, Evolution of non-terrestrial networks from 5g to 6g: A survey, IEEE Communications Surveys & Tutorials 24 (4) (2022) 2633–2672. doi:10.1109/comst.2022.3199901.

[173] T.-H. Nguyen, T. P. Truong, A.-T. Tran, N.-N. Dao, L. Park, S. Cho, Intelligent heterogeneous aerial edge computing for advanced 5g access, IEEE Transactions on Network Science and Engineering (2024). doi:10.1109/TNSE.2024.3371434.

[174] Starlink (2022).
URL https://www.starlink.com

[175] Kuiper (2022).
URL http://kuipersystemsgroup.com

[176] Oneweb (2022).
URL https://oneweb.net

[177] Y. Gao, J. Cao, P. Wang, J. Yin, M. He, M. Zhao, M. Peng, S. Hu, Y. Sun, J. Wang, S. Cheng, Y. Guo, Y. Du, Y. Cai, J. Huang, K. Qiu, Intelligent uav based flexible 5g emergency networks: Field trial and system level results, in: IEEE INFOCOM 2020 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), IEEE, 2020, pp. 138–143. doi:10.1109/infocomwkshps50562.2020.9162724.

[178] F. Volk, T. Schlichter, F. Kaltenberger, T. Heyn, G. Casati, R. T. Schwarz, A. Knopp, Field trial of a 5g non-terrestrial network using openairinterface, IEEE Open Journal of Vehicular Technology 3 (2022) 243–250. doi:10.1109/ojvt.2022.3175308.

[179] S. Kumar, A. K. Meshram, A. Astro, J. Querol, T. Schlichter, G. Casati, T. Heyn, F. Völk, R. T. Schwarz, A. Knopp, P. Marques, L. Pereira, R. Magueta, A. Kapovits, F. Kaltenberger, Openairinterface as a platform for 5g-ntn research and experimentation, in: 2022 IEEE Future Networks World Forum (FNWF), IEEE, 2022, pp. 500–506. doi:10.1109/fnwf55208.2022.00094.

[180] V. K. Shrivastava, S. Baek, Y. Baek, 5g evolution for multicast and broadcast services in 3gpp release 17, IEEE Communications Standards Magazine 6 (3) (2022) 70–76. doi:10.1109/mcomstd.0001.2100068.

[181] A. Rico-Alvarino, I. Bouazizi, M. Griot, P. Kadiri, L. Liu, T. Stockhammer, 3gpp rel-17 extensions for 5g media delivery, IEEE Transactions on Broadcasting 68 (2) (2022) 422–438. doi:10.1109/tbc.2022.3171508.

[182] 3GPP, Study on architectural enhancements for 5G multicast-broadcast services, Technical Report 3GPP TR 23.757, 3rd Generation Partnership Project (Mar. 2021).

[183] Y. Mao, C. You, J. Zhang, K. Huang, K. B. Letaief, A survey on mobile edge computing: The communication perspective, IEEE Communications Surveys & Tutorials 19 (4) (2017) 2322–2358. doi:10.1109/comst.2017.2745201.

[184] Y. Liu, M. Peng, G. Shou, Y. Chen, S. Chen, Toward edge intelligence: Multiaccess edge computing for 5g and internet of things, IEEE Internet of Things Journal 7 (8) (2020) 6722–6747. doi:10.1109/jiot.2020.3004500.

[185] 3GPP, Architecture for enabling edge applications, Technical Report 3GPP TR 23.558, 3rd Generation Partnership Project (Sep. 2022).

[186] 3GPP, Study on enhancement of support for edge computing in 5G core network (5GC), Technical Report 3GPP TR 23.748, 3rd Generation Partnership Project (Dec. 2020).

[187] 3GPP, Study on security aspects of enhancement of support for edge computing in the 5G core (5GC), Technical Report 3GPP TR 33.839, 3rd Generation Partnership Project (Mar. 2022).

[188] 3GPP, Study on 5G media streaming extensions for edge processing, Technical Report 3GPP TR 26.803, 3rd Generation Partnership Project (June 2021).

[189] 3GPP, Study on enhancements of edge computing management, Technical Report 3GPP TR 28.814, 3rd Generation Partnership Project (Sep. 2021).

[190] X. Foukas, G. Patounas, A. Elmokashfi, M. K. Marina, Network slicing in 5g: Survey and challenges, IEEE Communications Magazine 55 (5) (2017) 94–100. doi:10.1109/mcom.2017.1600951.

[191] S. Zhang, An overview of network slicing for 5g, IEEE Wireless Communications 26 (3) (2019) 111–117. doi:10.1109/mwc.2019.1800234.

[192] S. E. Elayoubi, S. B. Jemaa, Z. Altman, A. Galindo-Serrano, 5g ran slicing for verticals: Enablers and challenges, IEEE Communications Magazine 57 (1) (2019) 28–34. doi:10.1109/mcom.2018.1701319.

[193] T. C. Chuah, Y. L. Lee, Intelligent ran slicing for broadband access in the 5g and big data era, IEEE Communications Magazine 58 (8) (2020) 69–75. doi:10.1109/mcom.001.2000013.

[194] M. Maule, J. Vardakas, C. Verikoukis, 5g ran slicing: Dynamic single tenant radio resource orchestration for embb traffic within a multi-slice scenario, IEEE Communications Magazine 59 (3) (2021) 110–116. doi:10.1109/mcom.001.2000770.

[195] H. D. R. Albonda, J. Perez-Romero, An efficient ran slicing strategy for a heterogeneous network with embb and v2x services, IEEE Access 7 (2019) 44771–44782. doi:10.1109/access.2019.2908306.

[196] 3GPP, Study on enhancement of radio access network (RAN) slicing, Technical Report 3GPP TR 38.832, 3rd Generation Partnership Project (Jun. 2021).

[197] 3GPP TSG RAN Chair, Summary of RAN Rel-18 workshop, Electronic Meeting RWS-210659, 3rd Generation Partnership Project (Jul. 2021).

[198] W. Chen, J. Montojo, J. Lee, M. Shafi, Y. Kim, The standardization of 5g-advanced in 3gpp, IEEE Communications Magazine 60 (11) (2022) 98–104. doi:10.1109/mcom.005.2200074.

[199] X. Lin, An overview of 5g advanced evolution in 3gpp release 18, IEEE Communications Standards Magazine 6 (3) (2022) 77–83. doi:10.1109/mcomstd.0001.2200001.

[200] 3GPP, Artificial intelligence (AI)/machine learning (ML) for NG-RAN, 3GPP TSG RAN Meeting #94e RP-213602, 3rd Generation Partnership Project (Dec. 2021).

[201] 3GPP, Study on artificial intelligence (AI)/machine learning (ML) for NR air interface, Technical Report 3GPP TR 38.843, 3rd Generation Partnership Project (Dec. 2023).

[202] 3GPP, Study on network energy savings, 3GPP TSG RAN Meeting #94e RP-213554, 3rd Generation Partnership Project (Dec. 2021).

[203] H. Zhou, Y. Deng, L. Feltrin, A. Hoglund, Analyzing novel grant-based and grant-free access schemes for small data transmission, IEEE Transactions on Communications 70 (4) (2022) 2805–2819. doi:10.1109/tcomm.2022.3150787.

[204] 3GPP, Mobile terminated-small data transmission (MT-SDT) for NR, 3GPP TSG RAN Meeting #94e RP-213583, 3rd Generation Partnership Project (Dec. 2021).

[205] Y. Liu, H.-N. Dai, Q. Wang, M. K. Shukla, M. Imran, Unmanned aerial vehicle for internet of everything: Opportunities and challenges, Computer Communications 155 (2020) 66–83. doi:10.1016/j.comcom.2020.03.017.

[206] S. D. Muruganathan, X. Lin, H.-L. Maattanen, J. Sedin, Z. Zou, W. A. Hapsari, S. Yasukawa, An overview of 3gpp release-15 study on enhanced lte support for connected drones, IEEE Communications Standards Magazine 5 (4) (2021) 140–146. doi:10.1109/mcomstd.0001.1900021.

[207] 3GPP, NR support for UAV, 3GPP TSG RAN Meeting #94e RP-213600, 3rd Generation Partnership Project (Dec. 2021).

[208] 3GPP, Study on low-power wake-up signal and receiver for NR, 3GPP TSG RAN Meeting #94e RP-213645, 3rd Generation Partnership Project (Dec. 2021).

[209] 3GPP, Enhancement of NR dynamic spectrum sharing (DSS), 3GPP TSG RAN Meeting #94e RP-213575, 3rd Generation Partnership Project (Dec. 2021).

[210] RAN-Release 19 workshop - Meeting information 3GPP (2023).
URL https://portal.3gpp.org/Home.aspx#/55930-meetings

[211] 3GPP, Study on AI/ML model transfer phase2, Technical Report 3GPP TR 22.876, 3rd Generation Partnership Project (Jun. 2023).

[212] Intel, Mobility enhancements for Rel-19, 3GPP TSG RAN Rel-19 workshop RWS-230081, 3rd Generation Partnership Project, Taipei (Jun. 2023).

[213] NTT DoCoMo, AI/ML for air interface, 3GPP TSG RAN Rel-19 workshop RWS-230256, 3rd Generation Partnership Project, Taipei (Jun. 2023).
[214] LGU+, Views on AI/ML for NR air interface, 3GPP TSG RAN Rel-19 workshop RWS-230037, 3rd Generation Partnership Project, Taipei (Jun. 2023).
[215] Vivo, Views on Rel-19 AI/ML for air interface, 3GPP TSG RAN Rel-19 workshop RWS-230063, 3rd Generation Partnership Project, Taipei (Jun. 2023).
[216] CMCC, Further enhancements of SONMDT in Rel-19, 3GPP TSG RAN Rel-19 workshop RWS-230429, 3rd Generation Partnership Project, Taipei (Jun. 2023).
[217] 3GPP, Feasibility study on integrated sensing and communication (Rel-19), Technical Report 3GPP TR 22.837, 3rd Generation Partnership Project (Jun. 2023).
[218] Electronic Meeting, Motivation for integrated sensing and communication, 3GPP SA Meeting #95e SP-22084, 3rd Generation Partnership Project (Mar. 2022).
[219] Xiaomi, Motivation for integrated sensing and communication, 3GPP TSG RAN Rel-19 workshop RWS-230105, 3rd Generation Partnership Project, Taipei (Jun. 2023).
[220] 3GPP, Study on channel model for frequencies from 0.5 to 100 GHz, Technical Report 3GPP TR 38.901, 3rd Generation Partnership Project (Mar. 2021).
[221] Samsung, View on integrated sensing and communications, 3GPP TSG RAN Rel-19 workshop RWS-230221, 3rd Generation Partnership Project, Taipei (Jun. 2023).
[222] Electronic Meeting, Study on ambient power-enabled Internet of Things, 3GPP SA Meeting #95e SP-22085, 3rd Generation Partnership Project (Mar. 2022).
[223] Huawei, HiSilicon, Scope of ambient IoT work in Rel-19, 3GPP TSG RAN Rel-19 workshop RWS-230407, 3rd Generation Partnership Project, Taipei (Jun. 2023).
[224] Thales, Consideration on RAN1/2/3 led NTN topics for Release 19 (annex), 3GPP TSG RAN Rel-19 workshop RWS-230048, 3rd Generation Partnership Project, Taipei (Jun. 2023).
[225] 3GPP, Study on satellite access - phase 3, Technical Report 3GPP TR 22.865, 3rd Generation Partnership Project (Jun. 2023).
[226] I. MediaTek, S. Gatehouse, Novamint, M. Rakuten, Telstra, Study on complementary reuse of terrestrial spectrum in satellite deploy, 3GPP TSG RAN Rel-19 workshop RWS-230110, 3rd Generation Partnership Project, Taipei (Jun. 2023).
[227] 3GPP RAN, Summary for RAN Rel-19 Package: RAN1/2/3-led, Approval RP-232745, 3GPP RAN Meeting #102, Meeting (Dec. 2023).
[228] 3GPP highlights, issue 04 (May 2022).
URL `https://www.3gpp.org/ftp/Information/Highlights/2022_Issue04/mobile/3GPP_Highlights_I4.pdf`

Author Biography

Nhu-Ngoc Dao is an Assistant Professor at the Department of Computer Science and Engineering, Sejong University, Seoul, Republic of Korea. He received his M.S. and Ph.D. degrees in computer science at the School of Computer Science and Engineering, Chung-Ang University, Seoul, Republic of Korea, in 2016 and 2019, respectively. He received the B.S. degree in electronics and telecommunications from the Posts and Telecommunications Institute of Technology, Hanoi, Viet Nam, in 2009. Prior to joining Sejong University, he was a visiting researcher at the University of Newcastle, NSW, Australia, in 2019 and a postdoc researcher at the Institute of Computer Science, University of Bern, Switzerland, from 2019 to 2020. He is currently an Editor of the *Scientific Reports* and *PLOS ONE* journals. His research interests include network softwarization, mobile cloudization, intelligent systems, and the Intelligence of Things. Dr. Dao is a Senior Member of IEEE and a Professional Member of ACM.

Ngo Hoang Tu received the B.S. degree in Computer Networking and Data Communications from Ho Chi Minh City University of Transport (UT-HCMC), Vietnam, in 2020, and the M.S. degree with the Department of Smart Energy System Engineering at Seoul National University of Science and Technology (SeoulTech), South Korea, in 2022. He is currently pursuing the Ph.D. degree with the Department of Electrical and Information Engineering, SeoulTech. From January 2019 to January 2020, he worked as an Assistant Researcher with the Wireless Communication Laboratory, Posts and Telecommunications Institute of Technology (PTIT), Vietnam. From February 2020 to August 2020, he was a Lecturer with the Department of Computer Engineering, UT-HCMC. His research interests include wireless communications for MIMO systems, short-packet uRLLCs, intelligent reflecting surfaces, 6G infrastructures, and applied machine learning. He serves as an Editor for the *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems*.

Trong-Dai Hoang received the B.S. degree (Honor Program) in control engineering and automation from the Ho Chi Minh City University of Technology, Ho Chi Minh City, Vietnam, in 2019, and the M.S. degree in electrical and information engineering from Seoul National University of Science and Technology (SeoulTech), Seoul, Korea, in 2021. He is currently pursuing a Ph.D. degree in the School of Electrical and Data Engineering at the University of Technology Sydney, Australia. From September 2021 to July 2022, he worked as a Research Scientist at the Research Center for Electrical and Information Technology, SeoulTech. His research interests include the areas of applied machine learning, wireless communication, signal processing, and optimization.

Tri-Hai Nguyen received the B.S. degree (Honor Program) in computer science from the University of Information Technology, VNU-HCM, Ho Chi Minh City, Vietnam, in 2015, the M.Eng. degree in information and communication technology from Soongsil University, Seoul, South Korea, in 2017, and the Ph.D. degree in computer science and engineering from Chung-Ang University, Seoul, in 2022. He is currently a Research Professor with the Department of Computer Science and Engineering, Seoul National University of Science and Technology (SeoulTech), Seoul. His research interests include the Intelligence of Things, aerial computing, and beyond 5G/6G networks.

Luong Vuong Nguyen has been a Lecturer and Researcher at the Department of Artificial Intelligence, FPT University, Da Nang, Viet Nam, since October 2022. He has also been an Adjunct Professor at the Department of Artificial Intelligence, The Catholic University of Korea, Bucheon, South Korea, since November 2022. He received his Ph.D. degree in Computer Science and Engineering from Chung-Ang University, Seoul, Korea, in 2022; received an M.S. degree in the Department of Computer Engineering from The Da Nang University of Technology in Vietnam in April 2013; and received his B.S. in the Department of Mathematics from Da Nang University of Education and Science, Vietnam in July 2009. His research topics include knowledge engineering in data science by using data mining, machine learning, ambient intelligence, and logical reasoning.

Kyungchun Lee is a Professor with the Department of Electrical and Information Engineering, Seoul National University of Science and Technology (SeoulTech), Seoul, Republic of Korea. He received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, in 2000, 2002, and 2007, respectively. From April 2007 until June 2008, he was a Post-Doctoral Researcher with the University of Southampton, U.K. From July 2008 to August 2010, he was with Samsung Electronics, Suwon, South Korea. Since September 2010, he has been with SeoulTech, South Korea. In 2017, he was a visiting Assistant Professor with North Carolina State University, Raleigh, NC, USA. His research interests

include wireless communications and applied machine learning. He received the Best Paper Awards at the IEEE International Conference on Communications (ICC) and IEEE Wireless Communications and Networking Conference (WCNC) in 2009 and 2020, respectively.

Laihyuk Park received the B.S., M.S., and Ph.D. degrees in computer science and engineering from Chung-Ang University, Seoul, Korea, in 2008, 2010, and 2017, respectively. From 2011 to 2016, he was a Research Engineer with Innowireless, Bundang, Korea. From 2018 to 2019, he held the position of Assistant Professor at Chung-Ang University. He is an Assistant Professor with the Department of Computer Science and Engineering, Seoul National University of Science and Technology (SeoulTech), Seoul, Korea. His research interests include demand response, smart grids, and the Internet of Things.

Woongsoo Na received the B.S., M.S., and Ph.D. degrees in computer science and engineering from Chung-Ang University, Seoul, South Korea, in 2010, 2012, and 2017, respectively. He is currently an Assistant Professor with the Division of Computer Science and Engineering, Kongju National University, Cheonan, South Korea. Prior to joining Kongju National University, he was an Adjunct Professor with the School of Information Technology, Sungshin Womens University, Seoul, South Korea, from 2017 to 2018, and a Senior Researcher with Electronics and Telecommunications Research Institute, Daejeon, South Korea, from 2018 to 2019. His current research interests include mobile edge computing, flying ad hoc networks, wireless mobile networks, and beyond 5G.

Sungrae Cho is a Professor with the School of Computer Science and Engineering, Chung-Ang University (CAU), Seoul. Prior to joining CAU, he was an assistant professor with the Department of Computer Sciences, Georgia Southern University, Statesboro, GA, USA, from 2003 to 2006, and a senior member of technical staff with the Samsung Advanced Institute of Technology (SAIT), Kiheung, South Korea, in 2003. From 1994 to 1996, he was a research staff member with the Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea. From 2012 to 2013, he held a visiting professorship with the National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA. He received the B.S. and M.S. degrees in electronics engineering from Korea University, Seoul, South Korea, in 1992 and 1994, respectively, and the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2002. His current research interests include wireless networking, ubiquitous computing, and ICT convergence. He has been a subject editor of IET Electronics Letter since 2018 and was an area editor of Ad Hoc Networks Journal (Elsevier) from 2012 to 2017. He has served numerous international conferences as an organizing committee chair, such as IEEE SECON, ICOIN, ICTC, ICUFN, TridentCom, and the IEEE MASS, and as a program committee member, such as IEEE ICC, GLOBECOM, VTC, MobiApps, SENSORNETS, and WINSYS.