5 Software-Defined Air Interface (SDAI) for a Greener Network

Air interface is the most fundamental aspect of the physical layer, and has been continuously evolving in each wireless communication generation. The paradigm of air interface design in the previous four generations took peak data rate and system capacity as the dominant objectives and adopted a one-size-fits-all approach. As a result, a global optimized or trade-off air interface design, being not necessarily optimal for each individual application scenario, was adopted by previous standards. As for 5G, air interface design is expected to expand and support diverse use case scenarios and applications that will continue beyond the current 4G standards. Three typical use case scenarios for 5G have been identified: enhanced mobile broadband (eMBB), ultra-reliable low-latency communication (URLLC), and massive machine type communication (mMTC). eMBB aims to provide high data rate mobile broadband services; URLLC is designed for applications that have stringent latency and reliability requirements; and mMTC is the basis for ultra dense connectivity in internet of things (IoT). The consensus has been made in the community that 5G new radio (NR) should be more agile, efficient, and even able to update itself on demand. A unified air interface design to accommodate different applications and services shall become the major consideration in 5G NR and beyond.

This chapter focuses on the soft and green design of 5G NR air interface. It starts with an introduction of SDAI's framework, which is proposed by the CMCC. Then the wireless propagation channels are discussed, which serve as the foundation for designing a flexible air interface. Then the design considerations of frame structure, multiple input multiple output (MIMO), waveforms, multiple access (MA) scheme, full duplex, and signaling/control/protocol are elaborated.

5.1 SDAI Framework

SDAI extends the concept of "soft design" to the air interface in communication networks [1, 2]. Instead of a global optimized air interface, which is the trade-off among many factors, SDAI is highly motivated to meet the massive connections and diverse demands by reconfiguration and combination of multiple physical-layer building blocks, including frame structure, duplex mode, waveforms, MA scheme, modulation and coding, a MIMO transmission scheme, etc., as shown in Fig. 5.1. To enable SDAI, each potential building block could be predefined, then suitable building blocks are selected

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

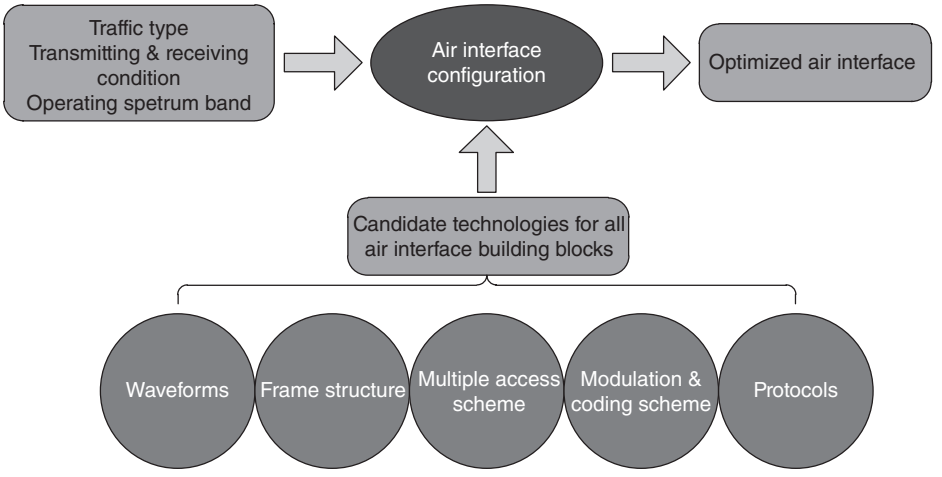


Figure 5.1 Building blocks of SDAI.

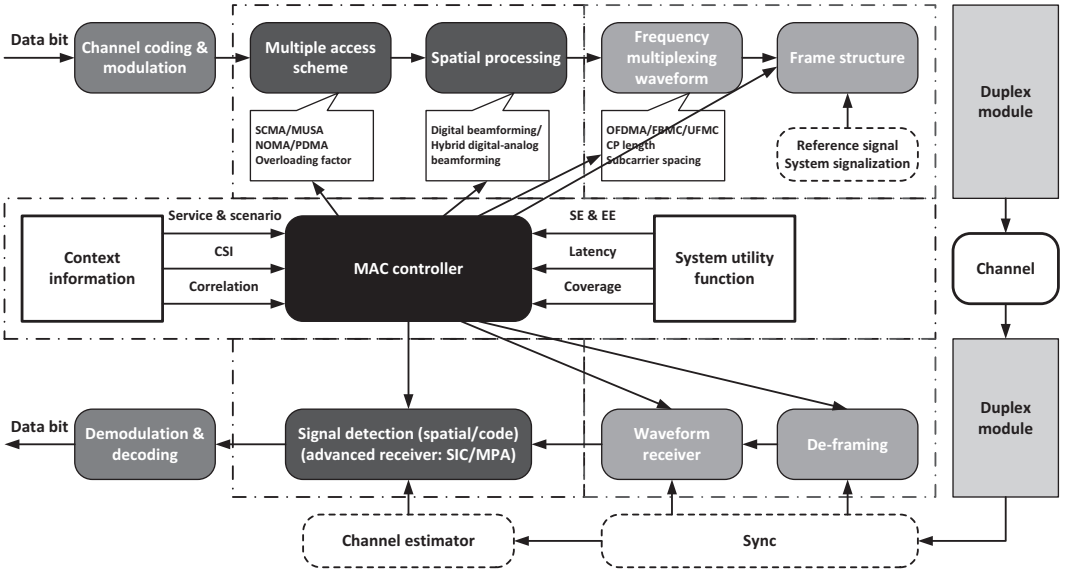


Figure 5.2 An illustrative structure of SDAI.

and properly configured and combined according to varying service requirements and network/UE capabilities to obtain the optimized technical solutions. It is expected to greatly improve the efficiency of wireless resources, reduce network deployment costs, and effectively cope with conceivable new scenarios and services. SDAI is expected to be a key driver of a green and soft RAN in 5G.

As shown in Fig. 5.2, the core of SDAI structure is an intelligent central controller. The controller is able to receive, sense, even predict the context information of the

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

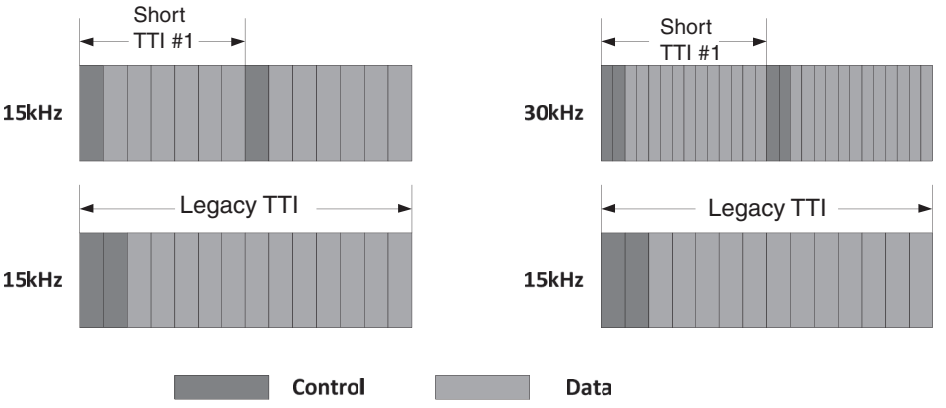


Figure 5.3 An illustrative example of short TTI.

network, including service types and requirements, traffic volume, channel state information, etc. Given the goal of optimizing specific system utility functions like spectrum efficiency, energy efficiency (EE), latency, coverage, and user QoE, the controller could select and configure the building blocks.

For example, a stringent user plane latency requirement of 1ms has been set for URLLC service in 5G standardization. Legacy frame structure and transmission time interval (TTI) length in LTE cannot meet the latency demand. A possible solution is to define a shorter TTI length in the air interface. A short TTI either contains fewer OFDM symbols than a legacy TTI, or contains same number of OFDM symbols but with a larger subcarrier spacing. The duration of a short TTI could be freely configured via the number of OFDM symbols and/or subcarrier spacing, as illustrated in Fig. 5.3. For constant URLLC traffic, the controller can configure legacy TTI transmission and short TTI transmission on separate frequency bands. Then filter-based OFDM waveform could be applied to mitigate the cross-numerology interference. For bursty URLLC traffic, the controller can configure both the legacy and short TTI transmissions on the same band to improve spectrum efficiency. The location of the short TTI transmission could be dynamically configured within the TTI duration.

Since there are diverse deployment scenarios and use cases emerging, or soon to emerge in future 5G networks, it is very important and beneficial for operators to deploy one network to support all use scenarios and use cases. To implement this, it is critical to adopt one unified and flexible air interface framework like SDAI to meet the diverse requirements of diversified usage scenarios. Following the E2E network slicing concept discussed in previous chapters, for each scenario, the air interface can be configured based on a physical layer (PHY) slicing, and corresponding layer 2 and layer 3 slicing. Each air interface slicing is tailored for specific service requirements and network/UE capabilities, while the coexistence of multiple slicing in one carrier needs to be well studied in a radio access network (RAN). In detail, each air interface slice may include the following aspects:

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

- Frame structure: Different DL and UL configurations and numerologies, variable subframe lengths to support different use cases. Flexible frame structure design is elaborated in Section 5.3.
- MIMO: Various single-user and multiple-user MIMO modes, with digital beamforming, analog beamforming, or hybrid analog and digital beamforming structures. Flexible MIMO design will be elaborated in Section 5.4.
- Waveforms: Various waveforms, including OFDM (orthogonal frequency division multiplexing), f-OFDM (filter-based OFDM), UFMC (universal filter multiple carrier), FBMC (filer bank multiple carrier), GFDM (generalized frequency division multiplexing), and OTFS (orthogonal time frequency spacing). Flexible waveform design will be elaborated in Section 5.5.
- MA schemes: Various orthogonal schemes and non-orthogonal schemes, such as MUSA (multiuser shared access), NoMA (non-orthogonal multiple access), SCMA (sparse code multiple access), and PDMA (pattern division multiple access). Flexible MA schemes will be elaborated in Section 5.6.
- Duplex: TDD, FDD, flexible duplex, full duplex, which will be elaborated in Section 5.7.
- Signaling: Layer 2 and layer 3 signaling associated with MAC, RLC, PDCP and RRC procedures, and possibly with cross layer interaction between RAN and the application layer. This will be covered in Section 5.8.
- Coding and modulation: various coding and modulation schemes, e.g., LDPC codes or polar codes.

5.2 Wireless Propagation in 5G Use Cases

The vision for 5G has been fueled by the development of two core PHY technologies that fundamentally set 5G NR apart from previous radio access technologies (RAT), namely, millimeter Wave (mmWave) and massive multi-input multi-output (MIMO) for below 6GHz frequencies. They both put forth a different paradigm that breaks with many current understandings of signal processing, device manufacturing, and network design, but most importantly the wireless propagation.

The physical layer technologies involved in air interface are generally built on the understanding of the fundamental of wireless propogation. In this section, we take a focused look upon the wireless propagation, i.e., its importance, its modeling methodology, existing channel models for cellular communications, and finally some challenging future research directions in 5G.

5.2.1 The Importance of Propagation Channels

The aim of this section is to briefly describe the fundamentals of radio wave propagation between the transmitter and the receiver. The main channel characteristics include path loss, shadow fading, and small-scale fading, as well as channel variations in time, frequency, and angle. The propagation channels dominate the actual performance of

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

any practical system, since physical law, described by Shannon's capacity equation, dictates the amount of information that can be carried through the wireless media. To evaluate the performance of a wireless system, one could rely on software simulation, which is probably the most cost-effective and time-saving evaluation method. It offers a mathematically tractable process and can be used to predict trends and average performance reasonably and accurately. Wireless channel models are generally needed for these simulations. Standardized channel models are furthermore essential to enable fair comparisons of different system proposals. Channel behaviors are well understood at traditional cellular frequency bands. However, in the realm of 5G, we see high-frequency bands (i.e., mmWave bands with wide bandwidth), and high mobility scenarios (e.g., 500km/hr for high-speed railway [HSR]) are considered for mobile access. Yet in both areas there is very limited research done for channel modeling. This hinges on the system design for the new technologies and new use cases.

5.2.2 Channel Modeling Principle and Fundamentals

Channel modeling should always be based on the understanding of the physics of the propagation. It needs to be an accurate representation of what the transmitter or the receiver "sees." Moreover, its mathematical formulation should be intuitive. Unlike propagation channels, channel models are dependent on the system in which it operates. This means it should only be as complex as necessary, neglecting any effects that won't affect the system performance.

Channel models used to be only in two dimension, however it is increasingly obvious that the elevation domain and the azimuth domain both have huge impacts on radio propagation, hence wireless performance. There are four basic propagation mechanisms: reflection, diffraction, scattering, and transmission (penetration), as illustrated in Fig. 5.4.

Propagation in the cellular and WLANs scenarios is more complex. In many cases, the receiver is placed in non-line-of-sight (NLOS) of the base station (BS) or access point (AP). The transmitted signal usually reflects off surfaces, diffracts around object edges, or transmits through obstacles, following different paths before reaching the receiver. This results in the multipath effect, where reflections, diffractions, and transmissions all attenuate the signal power. Each multipath signal can be abstracted as a vector, which comprises both magnitude and phase. After arriving at the receiver, vector summation occurs, leading to small-scale fading caused by the moving terminals and Doppler, which leads to frequency dispersion and time-selective fading.

5.2.3 Channel Modeling Methods in Cellular Systems

The state-of-the-art channel models can be classified as physical and analytical channel models, as shown in Fig. 5.5. The physical channel models concern the physical propagation environment, thus modeling signal parameters such as AOA, AOD, complex power, and time of flight. They can be further divided into deterministic channel models (e.g., ray-tracing and measurement-based models), geometry-based

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

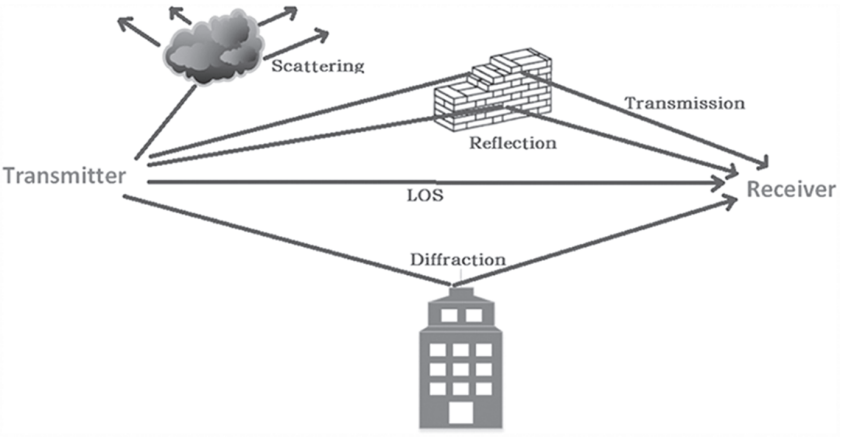


Figure 5.4 Wireless propagation.

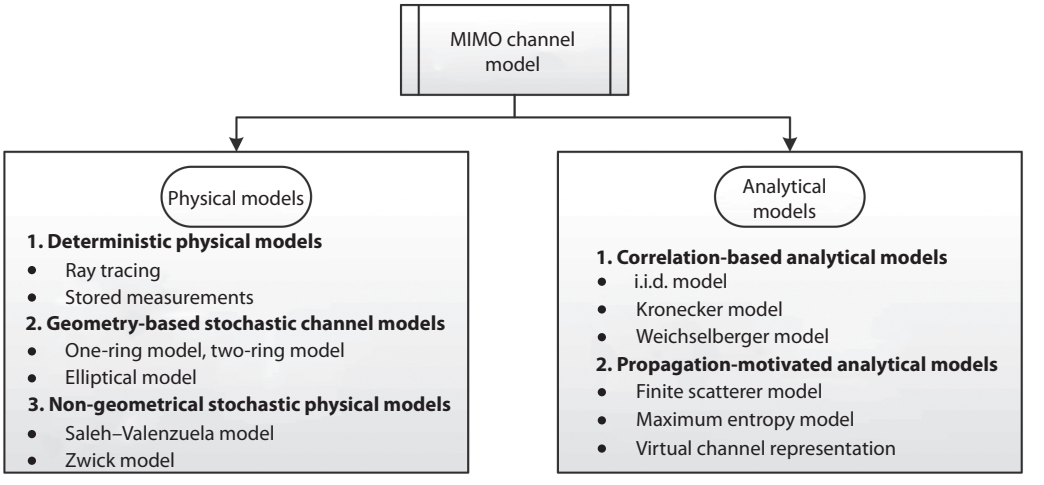


Figure 5.5 Traditional channel model classification.

(e.g., one/two-ring and elliptical models), and non-geometry stochastic channel models (e.g., Saleh-Valenzuela and Zwick models). Physical channel models can be in 2D or 3D.

The analytical channel models are derived from the statistical characteristics of the channels, which are obtained from mathematical representations such as the channel impulse response (CIR) between the transmitter and the receiver. Fitted models are often derived from the interpolation of measurement data points for a specific parameter. The analytical channel models can be divided into correlation-based stochastic models (CBSM) and propagation-based models (e.g., virtual channel representation [VCR]). Many of the current standardized wireless models can be put into these two categories, or a combination of them [3]. There is much great literature on the traditional channel modeling methodology for cellular systems, for instance [4–7]. Therefore, we will only focus on the new and challenging channel model research emerging in 5G.

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

5.2.4 New and Exciting Challenges in Channel Modeling

Massive MIMO Channels

Massive MIMO, is a disruptive 5G technology where the number of BS antennas grows to the hundreds, and aggressive spatial multiplexing supports tens of users with the same frequency and time radio resources [8]. Massive MIMO in the sub-6GHz band relies on the favorable propagation condition to deliver the superior spectral and energy efficiency. There is also the channel hardening effect, where the large number of BS antennas effectively average out the frequency selectiveness of users' channels due to spatial diversity [9, 10]. It is nevertheless very difficult to model the nonstationary phenomenon across the large antenna array elements and spherical wave effects when the UE is close enough to the BS and no longer in the far field of the array. The spatial, temporal, and angular correlations of the channel thus depart from the traditional understanding of MIMO channels [11].

There are already many research efforts in developing a massive MIMO channel model, for instance, the extension of the COST 2100 channel model [12], the extension of 3GPP-SCM [13], and the WINNER+ model [14], as well as mathematical models based on correlation or mutual coupling or geometry. They all have their accomplishment compared to traditional smaller MIMO models. However, none is without ambiguous assumptions or simply over-prediction of theoretical performance. Channel measurement campaigns are still needed to understand the necessary characteristics for an accurate model and its parameters.

Massive MIMO is not a stand-alone technology, and it is envisioned to be applied in other 5G applications, for instance, mmWave communication and distributed cellular deployment, as well as IoT scenarios. This further necessitates the understanding of the impact of large antenna arrays/RF chains on channel response seen by the BS and the UE before incorporating with other technologies.

mmWave Channels

Channel measurement is a necessary exercise for wireless researchers to understand the characteristics of a new spectrum. The mmWave band that is currently under consideration for mobile communications has a range from 30GHz to 100GHz. Since 2011, there had been many channel measurement campaigns conducted at 28, 38, 60, 72, and 73GHz. Channel modeling in this regard is critical for evaluating wireless technology in a timely and cost-effective manner. Considering many existing channel models and their evolution, such as the COST 2100 channel model, COST IC 1004 [15], ETSI model [16], and 3GPP TR 38.900 [17], it is always preferable to adapt and reuse the existing model structure. However, it is no longer enough to model the mmWave channel the same way as the microwave channel. Setting aside the obvious weather and oxgen absorption phenomenon, both large-scale and small-scale parameters need to be measured and analyzed. There has been continuous effort on modeling the large-scale spatial and angular parameters of the measured channels (e.g., path loss, shadow fading, delay spread, and angular power spectrum in azimuth/elevation domain) for standardization purposes in urban macro, urban micro, and indoor scenarios [18]. Due to lack

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

of measurement data, small-scale parameterization has not been validated extensively. Therefore, the current channel models such as the 3GPP 3D model, the QuaDRiGa model [19], the IEEE 802.11ad model [20], the MiWEBA model [21], the METIS model [22] and the mmMAGIC model [23] are not adequate to predict actual system performance.

Furthermore, there are unique mmWave channel characteristics that are yet to be specified and accurately modeled. For example, temporal channel statistics (e.g., the birth and death of multipath components [MPCs]) should be measured and investigated; the small-scale dynamics, e.g., diffuse scattering, significantly impact system performance and should be truthfully modeled; an intra-cluster model on PDP and power angular spectrum (PAS) allows for more accurate correlation modeling between MPCs; spatial consistency characteristics at such high frequency determine how the MIMO channel works; the blockage model of clusters is based on AOAs and attenuations, and its ability to represent the dynamic reality is highly desired. A first-attempt statistical channel model was presented in [24], taking the spatial, angular, and temporal statistics into consideration and consolidating them into a time cluster-spatial lobe (TCSL) approach.

It is increasingly obvious to us that, given the knowledge we have now about wireless propagation, there may not be a "one-size-fits-all" kind of channel model for such a vast range of spectrum and different deployment scenarios, as well as different levels of accuracy requirement.

Device-to-Device Channel

Mass connectivity has been a major selling point for 5G, where anyone and anything can be connected through the internet to facilitate everything from day-to-day convenience to high productivity. This requires device-to-device (D2D) communication, where the outdoor interaction with the environment is very different than traditional cellular device-to-infrastructure (D2I) channels (the indoor difference is less pronounced). The most obvious differences between D2D and D2I are the immediate surroundings of the transmitter, and the receivers being similar and correlated, and that both ends of the link could be mobile, i.e., dual mobility. There are also many more deployment scenarios than typically identified by standard bodies such as 3GPP; to name a few, indoor offices and shopping malls, as well as urban roads and highways. In each scenario, the link types, i.e., outdoor to outdoor (O2O), outdoor to indoor (O2I), indoor to indoor (I2I), or vehicle-to-vehicle (V2V), should also be taken into consideration when a more realistic channel representation is desired. Currently, there a three major established D2D channel models: WINNER, 3GPP, and COST 2100. However, the WINNER model derives the LSPs from deterministic maps, and the small-scale parameters do not have spatial and temporal consistency, leading to inaccurate MIMO channels; 3GPP D2D channels [25] are based on measurements, incorporating dual-mobility and Doppler effects, but making unrealistic assumptions about angular statistics such as uniform AOAs; COST 2100 is a GBSCM (geometry-based stochastic channel model) with cluster distribution based on real measurements and the visibility region feature enables moving clusters, yet it is not designed for D2D where the BS is not mobile.

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

In this section, we take V2V for illustrative purposes as a typical example of D2D channels. V2V considers both peer-to-peer 802.11p and future 5G-based communication as the data bearer. The final desired channel model should be general, easy to use, and able to transit smoothly among different scenarios.

Taking the V2V scenario at an intersection as an example, this wireless propagation parameter involves path loss, delay dispersion, Doppler spread, temporal variation, etc. Path loss is determined by the distances from TX/RX to side buildings, the width of the road, the distance between TX and RX, as well as the intersection spacing $[PL(d_r, d_t, w_r, x_t, i_s)]$. Delay spread can be modeled as random variables with a fitted distribution and a mean RMS value dependent on the deployment environment. Measurements have shown that the typical value is between 100 and 400ns. Temporal variation depends on the dynamics of the end nodes, since both ends of the link can move, sometimes at a fast speed, and the shadowing objects and scatters can move as well. The nonstationarities in the channel cannot be simply ignored, as what is done in 3GPP models; one of the less complicated solutions is GBSCM with randomly placed scatters or a tapped delay line model with a birth and death process and continuously changing delay. The final challenge is the ability to simulate multiple D2D links with the correct correlation/interaction between them. This is critical to the prediction of system performance under interference [26].

High-Speed Railway (HSR) Channels

High-speed railway is another much-discussed application that is considering 5G as a candidate technology for both safety-related critical data transmission and in-car passenger communication. Traditionally, GSM-R is used for trains at a low data rate of around 200kbps, but mainly for safety-related data. Currently, LTE-R and 5G-R are being actively researched and developed for next-generation railway communications [27]. It may have many new featured technologies, such as a distributed antenna system (DAS), coordinated multipoint (CoMP), or a mobile relay station (MRS). However, the challenges ahead will not be addressed properly if channel models are not accurate enough. There are unique features in HSR scenarios that lead to unique channel characteristics. Besides some challenges inherited from conventional trains such as high penetration losses, limited visibility in tunnels, and the harsh electromagnetic environment, typical hurdles involve fast handover, fast travel through diverse scenarios, large Doppler spreads, and nonstationarities. All the issues mentioned above are heavily dependent on the implementation method, i.e., mobile relay or direct link.

[28] has identified 12 different propagation environments, including viaduct, cutting, station, rural open area, and mountainous terrain, to name a few. Each environment has drastically different values and distributions of channel parameters. At microwave frequency, extensive measurements have been carried out and some practical models for power and fading are in use [29]; standardized models can be categorized as reused rural/urban models, such as a IMT-A channel model or simplified (quasi-LOS) model, neither of which is valid. At mmWave frequency, there is only one paper [30].

There is also a misunderstanding that "fortunately, most scenarios can be LOS, leading to simple channel characteristics." Unfortunately, this is a naive simplification of the

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

problem; due to rich scattering in some environments, the Ricean K factor is surprisingly very small, and even becomes negative in tunnels. Additionally, there is not just a Doppler shift, but a Doppler spread in most situations, even in LOS. Angular spread is also very pronounced across short distances, e.g., 20 degrees of RMS angular spread in 100m in viaduct (assuming the train speed is 500km/hr, the time it takes is 0.72s). Finally, the power delay profile is closely related to the immediate environment; as we can see in [31], the measured PDP has a strong periodicity due to the power poles along the tracks. Moving forward, nonstationary HST channel models could be deterministic or stochastic, or the hybrid of the two, catering to different applications of the channel models. Deterministic models can be pure ray-tracing and a random graph; stochastic models can be GBSCM (e.g., a finite-state Markov channel can effectively capture the dynamic nature of fast fading/time-varying in some environments). There is also the possibility that large-scale fading models and small-scale models are separately developed.

5.2.5 Concluding Remarks

This section began with a brief description of wireless propagation and the fundamental principles in channel modeling, as well as the characterization of different model methods. The main focus herein is identifying the channel modeling challenges and some possible directions/solutions in various 5G scenarios, i.e., massive MIMO, mmWave, D2D, and HST. In summary, channel modeling work is the foundation for the success of 5G, where diversified use cases and requirements demand statistically accurate and easily applicable channel models. Extensive measurements and innovative methods (e.g., leveraging wireless big data) are needed to blaze the challenges ahead. More importantly, given the fact that channel conditions vary significantly in diversified scenarios and extensive measurements and modeling methods are required, the 5G air interface design is strongly motivated to be as flexible and efficient as possible, and capable of configuring the physical-layer and higher-layer building blocks and parameters. In the following subsections, the soft and green design of physical-layer technologies will be elaborated, including frame structure, waveform, MA, MIMO, and duplex.

5.3 Flexible Frame Structure

Frame structure is the basic DL and UL operation framework for wireless communication systems, which specifies where and when the signaling, control, and data should be transmitted. In LTE, frame structure is generally designed with relatively fixed settings that consider the worst scenarios, which simplifies the system design but at the cost of efficiency. As diversified scenarios, such as eMBB, URLLC, mMTC, and the wide range of spectrum defined in 5G, instead of the traditional "one-size-fits-all" design, the frame structure in 5G is envisioned to be more agile and efficient. It should be dynamically configured to adapt to various propagation and application scenarios. On the other hand, facing new challenges posed by 5G, such as the extremely high mobility (up to

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

500km/hr), ultra-low latency (less than 0.5ms for URLLC), and massive connectivity, the frame structure is expected to be comprehensively designed to fulfill these stringent demands.

In this section, we begin with the frame structure design principles in SDAI for 5G. Then, key features of the flexible frame structure, e.g., scalable numerology, service multiplexing, configurable subframe/scheduling unit, flexible reference signal/scheduling, and HARQ timing, are discussed. Finally, the standardization progress of frame structure in 3GPP is also summarized, along with the future research directions on full duplex frame structure.

5.3.1 Frame Structure Design Principles

In order to address the great challenges in 5G and facilitate SDAI, the frame structure should be more flexible compared with that of 4G era. In the following, some basic design principles for frame structure are listed:

- Allowing for scalability to address different services requirements;
- Support for efficient multiplexing of different services, e.g., eMBB, URLLC, and mMTC;
- Support for an extremely wide range of physical properties, e.g., very wideband, narrow band, TDD, FDD, sub 6 GHz, and mmWave bands;
- Support for dynamic TDD assignment with efficient interference management;
- Support for self-contained subframe with a single interlace structure (ACK/NACK in the same subframe) and possible multiple interlace structure for forward compatibility;
- Support for tight coupling across aggregated carriers (i.e., supporting carrier aggregation).

With the above-mentioned principles in mind, some key features of the flexible frame structure are illustrated in the following part.

Scalable Numerology

To support a wide range of services, deployment scenarios, and spectrum, the numerologies (including subcarrier spacing, cyclic prefix length, and TTI length) of the frame structure need to be scalable and flexibly configurable.

For eMBB services, LTE numerology of 15kHz subcarrier spacing works well below 6 GHz. For mMTC narrow-band services, a narrower subcarrier spacing, e.g., 3.75KHz, is desired for higher capacity when considering the same coverage. For URLLC service, a larger subcarrier spacing, e.g., 60KHz, is preferred for latency reduction. Besides the service aspects, for numerology design, deployment-related attributes (such as carrier frequency, channel characteristics, inter-site distance, UE speeds, and possible transmission schemes) also should be taken into account. In addition, implementation cost and complexity are also crucial factors for the numerology design. With much broader bandwidth defined for 5G, the required FFT size and frequency domain signal processing with RE- or RB-level granularity may reach a point where

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

Table 5.1 Numerology for diverse services, deployments and spectrum.

Motivation	Scenarios	Subcarrier spacing	CP length	TTI length
Diverse service	eMBB	>=15kHz	Depends on scenarios	Depends on scenarios
	mMTC	<=15kHz	Longer CP	Longer TTI
	URLLC	FFS	Depends on scenarios	Shorter TTI
Diverse deployment	Low to medium UE speed	15kHz	Depends on services	Depends on services
	High UE speed	>=60kHz	Depends on services	Depends on services
Diverse spectrum	Sub-6GHz	Depends on scenarios	Depends on scenarios	Depends on scenarios
	Above 6GHz	Larger carrier spacing	Shorter CP	Shorter TTI

the implementation cost and complexity become unacceptable if LTE legacy subcarrier spacing is utilized, especially for the mmWave bands. Targeting a unified air interface for diversified services and deployment scenarios with all potential available spectrum, the scalable subcarrier spacing and TTI length could be the starting point.

In Table 5.1, some examples on the numerology design for diverse services, deployments, and spectrum are shown [32].

Service Multiplexing

To have an efficient utilization of scarce air-interface resources, multiplexing of transmissions with different latency and/or reliability requirements for eMBB/URLLC/mMTC should be supported in 5G. When UEs are scheduled in different networks or frequency bands, no attention needs to be paid for services multiplexing. But, for cases in which UEs with different services are scheduled in the same band and same network, flexible frame structure with variable numerologies are suggested to be considered and applied, so as to satisfy the requirements of different services simultaneously.

Different services may require different numerologies. The question of how to support multiplexing of different numerologies in the same frame structure needs be addressed carefully. Both time division multiplexing (TDM) and frequency division multiplexing

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

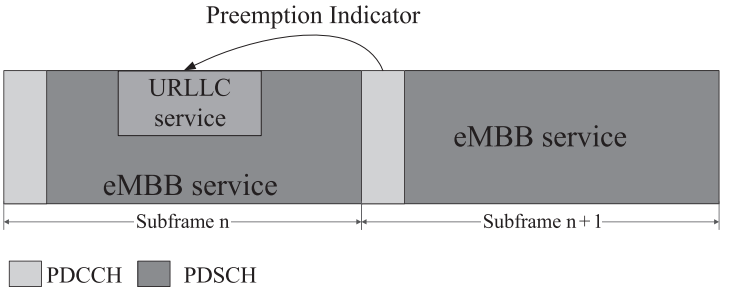


Figure 5.6 Multiplexing of URLLC and eMBB services.

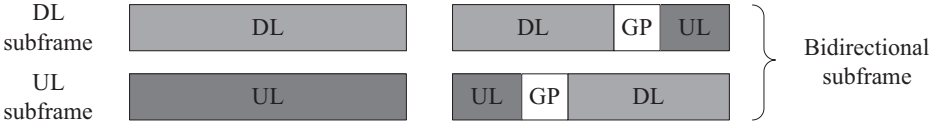


Figure 5.7 Three types of time domain structure.

(FDM) can be considered. For FDM, the interference of both UL to DL and DL to UL when considering the multiplexing of different services needs to be addressed. Applying TDM is in general more beneficial for multiplexing long and short scheduling frames. A typical use case for applying TDM is mission-critical applications. They are usually characterized by a bursty transmission and supported via time domain puncturing. For example, dynamic resource sharing between URLLC and eMBB can be supported by transmitting URLLC traffic in resources scheduled for ongoing eMBB traffic, known as puncturing or preempted transmission, as shown in Fig. 5.6. To avoid the severe performance loss of eMBB in this case, an indicator signal may be necessary to let the eMBB UE know the existence of URLLC transmission.

Configurable Subframe

In order to meet the requirements of different services, three types of subframes are recommended to be supported, i.e., DL subframe, UL subframe, and bidirectional subframe, as illustrated in Fig. 5.7.

Considering the requirement of URLLC on fast response to scheduling, transmission and feedback, bidirectional subframe seems to be a good solution to make self-scheduling and self-feedback possible. A bidirectional subframe contains a DL transmission region (containing DL control, RS and/or data), guard period (GP), and an UL transmission region (containing UL control, RS and/or data). The overhead of GP can be configured flexibly to cater to different scenarios.

Even though a bidirectional subframe provides adequate flexibility for DL and UL scheduling and timing, it is necessary to support a full DL and UL subframe, so as to reduce overhead and ensure coverage, especially in scenarios with low frequency and macro coverage. They are more efficient when deployed on a paired spectrum, and on an unpaired spectrum when latency is not a problem. In addition, it is suggested to keep as

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

many commonalities as possible for operating these subframe types on paired/unpaired spectrums and licensed/unlicensed spectrums.

In summary, three types of subframes should be supported in 5G: DL subframe, UL subframe, and bidirectional subframe. For different use cases, the flexible combination of these types of subframes can be considered.

Configurable Scheduling Unit

Different services are characterized by different features, for example, URLLC traffic may have small packet size but strict delay requirements, while eMBB service may ask for very high data rate with medium latency, and mMTC may put stringent requirements on coverage and connection numbers. It is possible that a common scheduling time unit is defined for all the services, but apparently it is not efficient. To satisfy the delay requirement of URLLC, the time duration of the common time unit should be short enough, and a short time unit will result in limited resource elements in it. For eMBB service with large packet size, several continuous time units may be needed to one UE schedule, with a common short time unit, as URLLC, the control channel overhead, and feedback overhead may increase linearly. Thus, it is more efficient to define a configurable scheduling time unit.

Different from the 1ms scheduling unit defined in LTE, for 5G, mini-slot-, slot-, and slot-aggregation-based scheduling should be supported. Note that mini-slot can be composed of fewer OFDM symbols. A subframe should contain an integral number of mini-slot for timing alignment. Mini-slot-based scheduling can be utilized for traffic with small packet size but strict delay requirements, e.g., URLLC services, while the slot aggregation scheduling unit is suit for the large packet size with stringent requirements on coverage. Data transmitted in the scheduling unit should be self-decodable with its control channel, reference signals, and A/N feedback. With configurable different scheduling time intervals, and with different packet sizes, delay and coverage requirements can be satisfied with high efficiency.

Flexible Reference Signal Design

In general, reference signals (RS) are used for data demodulation, phase tracking, time/frequency tracking, channel state information (CSI) measurement, radio link monitoring, RRM measurement, etc. Various RS signals, such as DMRS, SRS, CSI-RS, and PT-RS are being heatedly discussed in 3GPP NR. Basically, RS design needs to consider the tradeoff between the estimation performance and the overhead. For example, refarding the high-speed train scenario with 500km/hr mobility, denser DMRS is needed to combat the rapidly varying channel for accurate channel estimation. As for the low and medium mobility scenario, much sparse DMRS can be configured to reduce the DMRS overhead.

Flexible Scheduling/HARQ Timing

To support diverse services with different latency requirements and different UE processing capability, configurable scheduling and HARQ timing is expected to be sup-

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

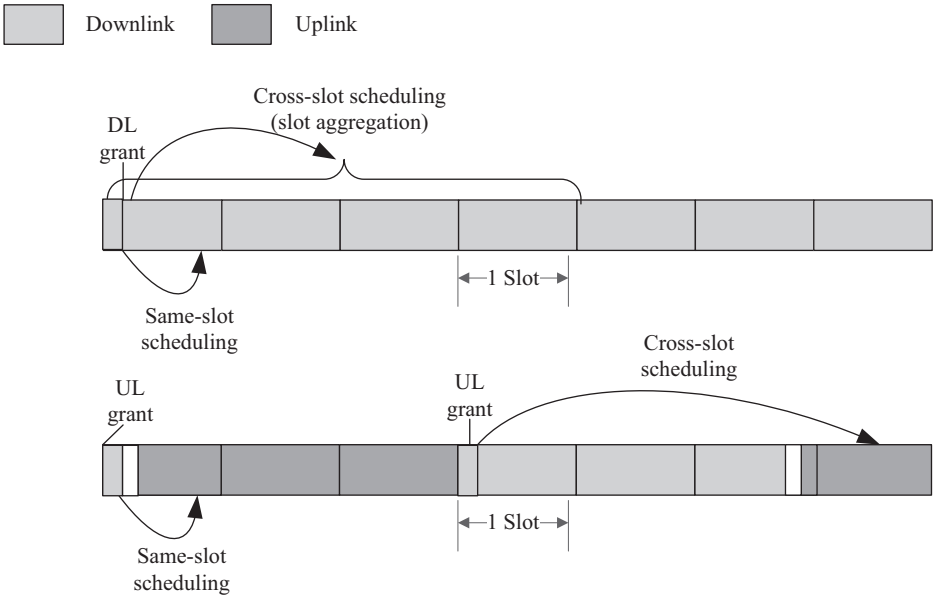


Figure 5.8 An example of configurable scheduling timing.

ported in 5G. Four types of timing relationship need to be considered for the scheduling and HARQ timing.

- K0: Delay between the DL grant and corresponding DL data (PDSCH) reception
- K1: Delay between DL data (PDSCH) reception and the corresponding acknowledgement transmission on UL
- K2: Delay between UL grant reception in DL and the UL data (PUSCH) transmission
- K3: Delay between the ACK/NAK reception in UL and the corresponding retransmission of data (PDSCH) on DL

An example of configurable scheduling timing is shown in Fig. 5.8.

For the DL/UL scheduling, at least the same-slot scheduling and the cross-slot scheduling should be supported. Note that the value of minimum timing depends on many factors such as transmission block size, data RE mapping, the RS location, channel coding and cell range, UE processing capability, etc. Scheduling multiple DL/UL slots from a single DL control occasion enables the reduction of both signaling overhead for NR data transmissions and the interference caused to neighboring cells.

Figure 5.9 illustrates an example of flexible HARQ timing. For low latency service or small packet transmission, self-contained properties can be supported including the short HARQ timing between DL data and the corresponding A/N, UL grant, and the corresponding UL data. For noncritical eMBB service or large packet transmission, more processing time is needed, and accordingly HARQ timing is longer.

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

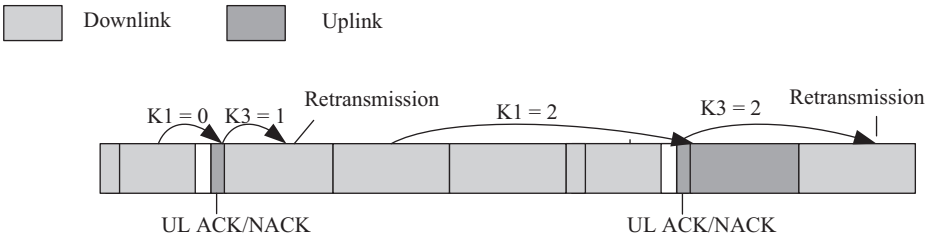


Figure 5.9 An example of flexible HARQ timing.

5.3.2 Progress of Frame Structure in 3GPP 5G NR

Two types of timing units, fixed length and variable length, are defined in state-of-the-art standardization on 5G frame structure. Timing units with fixed length include radio frame and subframe. A radio frame has a 10 ms duration and a subframe has a 1 ms duration. They are chosen to be consistent with LTE to strive for a better LTE-NR coexistence performance in case of co-site deployment. On the other hand, to support diverse services and multiple numerologies, timing units of varying length are also defined, including slot and mini-slot. A slot contains 14 OFDM symbols for normal CP length (nearly 7% CP overhead) and a mini-slot contains X OFDM symbols, where X is less than 14 and greater than 1. Mini-slot mainly targets fast scheduling service and/or delay-critical service, e.g., URLLC. Obviously, the length of a slot depends on the subcarrier spacing and that of a mini-slot depends on both subcarrier spacing and number of OFDM symbols. Note that, for extended CP length (nearly 25% CP overhead), a slot only contains 12 OFDM symbols. Currently, except for 60kHz subcarrier spacing, whether to support extended CP length on other subcarrier spacing is still undecided. The motivation to support 60kHz subcarrier spacing with extended CP length derives from the high-speed train scenario.

Slot format is another aspect in standardization on frame structure that shows flexibility. Now it has been agreed that "A slot can contain all downlink, all uplink, or at least one downlink part and at least one uplink part [54]", which exactly conforms to our principles on frame structure design in SDAI. Moreover, the slot format can be dynamically indicated to UEs through layer 1 signaling, rather than the only semi-static ways in LTE. This further provides adequate flexibility on DL/UL traffic adaption and scheduling.

5.3.3 Concluding Remarks

A flexible frame structure is the basis for green and soft RAN operation, which is capable of flexible configuration of UL and DL slots, numerology, scheduling unit, HARQ timing, RS patterns, etc. This is quite motivated in the 5G era to meet service requirements with diversified KPIs in various usage scenarios. Future research directions of frame structure may include the possible transition to full duplex when full duplex technologies become mature and ready to be implemented in cellular networks.

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

5.4 Flexible MIMO

MIMO techniques have been widely utilized in 4G LTE systems, where multiple MIMO schemes including diversity, spatial multiplexing, and MU-MIMO are specified [33]. These schemes are implemented in basedband via the digital beamforming structure. In 5G NR, a distinguishing configuration is that digital beamforming, analog beamforming, and hybrid beamforming will all possibly be considered in system deployment [34]. Different structures may be employed at both BS and UE in various scenarios and frequency bands. For example, analog beamforming may be more suitable for indoor scenario in mmWave bands. When more uses need to be supported in spatial domains, hybrid beamforming is motivated, where on top of analog beamforming, digital beamforming may further help to reduce the inter-user/inter-beam interferences. For lower frequency bands in 5G, traditional digital beamforming may be the most suitable. This necessitates the design of flexible MIMO with a unified beamforming architecture and a unified CSI acquisition/feedback mechanism. In addition, as the antenna number is expected to increase both in BS and UE in 5G communications, the energy efficiency will be an important performance indicator, necessitating energy efficient designs.

In this part, a unified MIMO framework is presented, which includes analog, digital, and hybrid beamforming as special cases. Typical hybrid beamforming structures are also investigated, with various beamforming algorithms surveyed. Furthermore, energy-efficient design considerations of hybrid beamforming structure are presented. Finally, the standardization of hybrid beamforming is discussed.

5.4.1 Unified Framework for MIMO Techniques for 5G

In the last 10 years, the evolution of MIMO techniques tends to employ a large number of antennas (usually hundreds of antennas) at the BS to serve tens of users in the same time and frequency resource, which is commonly referred to as massive MIMO [35]. Massive MIMO is not a straightforward extension of conventional small-scale MIMO. Its systems can acquire "channel hardening," such that the uncorrelated noises and channel vectors for different users are averaged out, and simple linear signal processing procedures can achieve near-optimal performance.

With the severe spectrum shortage in conventional cellular bands, massive MIMO in mmWave bands can potentially help to meet the anticipated demands of mobile traffic in 5G era. There are many challenging issues in the implementation of digital beamforming on mmWave, including complexity, energy consumption, cost, etc. In a practical deployment, hybrid beamforming structures can be an important alternative choice and have been proposed as an enabling technology for 5G cellular communications [36–39].

The main concept of hybrid beamforming is to divide the traditional baseband signal processing into digital and analog domains. A unified framework of hybrid beamforming is illustrated in Fig. 5.10, where transmit data from N_s ports are mapped onto a N_t^{RF} transmit and receive unit (TXRU) via digital beamforming, and further mapped onto N_t antennas via analog beamforming. The mapping PA (power amplifier) is determined by the connections between TXRU and antennas, and the phase and amplitude of each RF

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

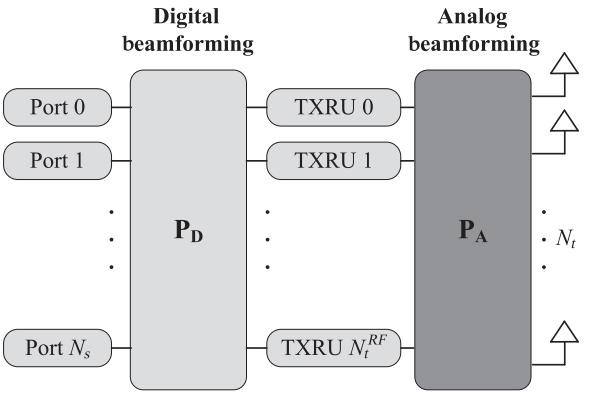


Figure 5.10 A block graph for hybrid beamforming.

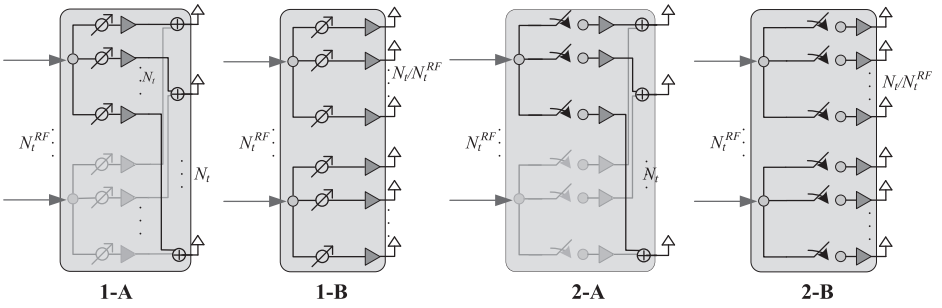


Figure 5.11 Typical structures of the analog part of the hybrid beamforming.

path. This architecture is a generic framework. For example, when $N_{\mathrm{t}}^{\mathrm{RF}} = N_{\mathrm{t}}$ and each TXRU is directly linked to each antenna element, this architecture would be a typical digital beamforming structure, which doesn't need analog processing. Based on this generic architecture, various specific realizations of hybrid beamforming architecture and corresponding beamforming schemes can be developed.

Figure 5.11 shows typical structures of the hybrid beamforming. Generally speaking, one category of hybrid beamforming architecture is the full-array architecture (as shown in Fig. 5.11 (1-A) and (2-A), in which each stream of the signal is transmitted via the whole antenna array [40]). Another kind of hybrid beamforming architecture is called the sub-array architecture [41] (as shown in Fig. 5.11 Architecture 1-B and Architecture 2-B), in which each stream of the signal is transmitted using a set of antenna elements instead of the whole antenna array. The key difference between full-array structures and sub-array structures is that the full-array structure totally needs $N_t^{RF} \times N_t$ analog devices (e.g., phase shifters or switches), since each RF chain is connected to all N_t antennas, while the sub-array structure only requires N_t analog devices, since each RF chain is only connected to N_t/N_t^{RF} antennas, leading to this leads to greatly increased complexity for the former. The difference between the structures with phase shifters and

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

those with switches is that the cost in the latter case is much reduced, though there can be some performance loss.

5.4.2 Schemes of Hybrid Beamforming

We consider a typical massive MIMO system with hybrid beamforming structure, where the BS with N_t transmit antennas sends N_s independent data streams to the user with N_r receiving antennas. Furthermore, it could be assumed that the BS and the user have N_t^{RF} and N_r^{RF} chains, respectively, which satisfy $N_s \leq N_t^{RF} \leq N_t$ and $N_s \leq N_r^{RF} \leq N_r$. In the hybrid precoding structure, as shown in Fig. 5.10, the hybrid precoding matrix $\mathbf{P} \in \mathbb{C}^{N_t \times N_s}$ at the BS can be written as the product of two parts: the first part is a low-dimension digital precoding matrix $\mathbf{P}_D \in \mathbb{C}^{N_t^{RF} \times N_s}$; the second part is a high-dimension analog precoding matrix $\mathbf{P}_A \in \mathbb{C}^{N_t \times N_t^{RF}}$, i.e., $\mathbf{P} = \mathbf{P}_A \times \mathbf{P}_D$, where $\mathbb{C}$ is the set of complex numbers. Note that these remarks on precoding matrix can also be applied to the combining matrix. Thus, the transmitted signal vector $\mathbf{x}$ is

$$\mathbf{x} = \mathbf{P}\mathbf{s} = \mathbf{P}_A\mathbf{P}_D\mathbf{s}. \tag{5.1}$$

where $\mathbf{s} \in \mathbb{C}^{N_s \times 1}$ denotes the source signal vector. After receiving the signal vector, the user utilizes a hybrid combining matrix $\mathbf{W}$ to combine the signal vector:

$$\mathbf{y} = \sqrt{\rho}\mathbf{W}^H\mathbf{H}\mathbf{P}_A\mathbf{P}_D\mathbf{s} + \mathbf{W}^H\mathbf{n}. \tag{5.2}$$

where ρ is the average received power, $\mathbf{H} \in \mathbb{C}^{N_r \times N_t}$ denotes the channel matrix between BS and the user, and $\mathbf{n} \in \mathbb{C}^{N_r \times 1}$ is the additive noise vector.

The optimal beamforming matrix could be inferred through an optimization method to achieve certain objectives such as maximizing sum-rate or minimizing power consumption with QoS constraints. Following the pioneering work [42] published in 2014, various hybrid beamforming schemes have been recently proposed to achieve different trade-offs between performance and costs, as surveyed in [43]. In [37], two alternating hybrid beamforming methods are proposed to jointly optimize the analog and digital beamforming matrices to maximize the achievable rate with different practical constraints. The authors in [44] propose to decompose the total achievable rate optimization problem with non-convex constraints into a series of simple subrate optimization problems, each of which only considers one subantenna array.

In the remainder of the subsection, some prior works are surveyed, which consider low-resolution and low-cost ADCs/DACs and phase shifters in the hybrid beamforming algorithm design. A DAC/ADC (digital-to-analog converter or analog-to-digital converter) performs the function of transforming the digital (analog) signal to the corresponding analog (digital) signal.

1. **Hybrid beamforming with few-bits ADC/DAC**: With bandwidths on the order of a gigahertz in mmWave communication systems, high-resolution ADCs or

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

DACs become a power consumption bottleneck and cause excessive signal processing. One solution is to employ low resolution one-bit or few-bits ADCs. Considering the low cost of low-resolution ADCs (or DACs), it would be a particularly attractive solution for massive MIMO systems.

Since the channel state information at transmitter (CSIT) and channel state information at receiver (CSIR) needs transmitter beamforming and receiver combining, the CSI estimation for one-bit (or few-bits) ADC is particularly important for massive MIMO hybrid beamforming. For conventional MIMO systems, the closed-form ML estimator of such a SISO channel has been derived in [45]. Yet, for ultra wideband mmWave systems, these works may not be efficient for their frequency-selective characteristics. One method is to transmit the burst reference signal and estimate the CSI of each tap of the channel separately. Another more efficient scheme is to employ the generalized approximate message passing method using the channel correlation information based on the sparsity of the mmWave channel. Since channel estimation errors with one-bit ADCs would decrease with the sparsity of the channel [46], it would be more convenient to use compressive sensing techniques for channel estimation with relatively few CSI measurements (one-bit or few-bit CSI).

As for transmitter beamforming, a detailed capacity analysis of one-bit quantized MIMO systems with available CSIT is provided in [47]. At low and medium SNRs, when CSIT is assumed to be available, simple channel inversion beamforming is shown to be nearly optimal if the channel has full row rank. At high SNRs, a specific beamforming scheme was proposed to achieve system capacity even if the channel matrix is full rank. Hence, this transmitter beamforming technique would eliminate the gap between unquantized and quantized CSI. [48] studied the impact of CSI on the sum capacity of massive MIMO systems with quantized hybrid beamforming where the RF beamformer is selected from a finite size codebook. Considering the sparsity characteristics of the mmWave channel, channel inversion beamforming may be better than eigen-beamforming. This means that, kinds of simplified hybrid beamforming optimizations would be suitable for one-bit ADCs. With the increasing of the number of antenna elements, CSI acquiring (or recovery) would also be more challenging. In [49], a method of hybrid beamforming was proposed to reduce the quantization errors introduced at the analog beamformer part, which would lead to performance degradation.

2. **Hybrid Beamforming with few-bits phase shifters:** Due to limitations in cost and power supply, analog beamforming with a low-resolution phase shifter, instead of pure baseband digital beamforming, tends to be more favorable in mmWave hybrid beamforming systems. In IEEE 802.11ad, beamforming is based on a codebook with a 2-bit phase shifter. However, the use of a low-resolution phase shifter would degrade the link performance, as analyzed in [50]. Moreover, the gain loss brought by the low-precision phase shifter is limited; a 3-bit phase shifter can get a performance close to the ideal one.

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

5.4.3 EE–SE Analysis of Hybrid Beamforming

The EE and SE analysis of digital and hybrid beamforming has been addressed in many previous works, e.g., [51–53]. Take the sub-array structure as an example, where perfect analog beamforming is assumed within each sub-array with M antennas, which points to one user (there are N users in total). Assuming there is no inter-user interference, i.e., there is proper user scheduling (the BS schedules users with orthogonal channels), then the sum capacity of this structure for N users is:

$$C = W \times N \times \log\left(1 + \frac{M\eta_{\mathrm{PA}}P}{WN_0}\right), \tag{5.3}$$

where W is the bandwidth, P is transmit power of each transceiver (the total power of M antenna PAs), η_{PA} is the PA efficiency, and N_0 is the thermal noise density. Without loss of generality, the channel gain is assumed to be the unity. The SE of this structure is:

$$\eta_{\mathrm{SE}} = C/W = N \times \log\left(1 + \frac{M\eta_{\mathrm{PA}}P}{WN_0}\right). \tag{5.4}$$

Because the accurate power model is nontrivial, the following simple power model is used:

$$P_{\mathrm{total}} = NP + P_{\mathrm{static}} = NP + NP_0 + P_{\mathrm{common}} + NMP_{\mathrm{rf\_circuit}}, \tag{5.5}$$

where P_{total} is the total power; NP is the RF power of N transceivers; P_{static} is the static power of the BS, including NP_0, which scales with N; P_{common}, which is common for any number of transceivers; and $NMP_{\mathrm{rf\_circuit}}$, which scales with NM. The relationship between EE and SE is

$$\begin{aligned}\eta_{\mathrm{EE}} &= C/P_{\mathrm{total}} \\ &= \frac{\eta_{\mathrm{SE}}}{\left(2^{\frac{\eta_{\mathrm{SE}}}{N}} - 1\right)\frac{N_0}{\eta_{\mathrm{PA}}}\frac{N}{M} + \frac{NP_0 + P_{\mathrm{common}} + NMP_{\mathrm{rf\_circuit}}}{W}}.\end{aligned} \tag{5.6}$$

Therefore, for a required SE, the hybrid LSAS beamforming should be designed to maximize EE through joint design of N, M, P_0, P_{common}, $P_{\mathrm{rf\_circuit}}$, and η_{PA}.

Relationship at Green Points

When we take circuit power into consideration, there is a "green" point on the EE–SE curve where EE is at its maximum and is denoted η_{EE}^{*} [36]. Here, we discuss two cases for the $N \times M$ sub-array hybrid beamforming structure: 1) $NM = L$ (i.e., the total number of antennas is fixed as L, but N and M are variable), and 2) N and M are independent. For the former case, we allow the first-order derivative of EE over SE to be zero:

$$\eta_{\mathrm{EE}}^{'} = \frac{aN^2\left(2^{\frac{\eta_{\mathrm{SE}}}{N}} - 1\right) + bN + c - \eta_{\mathrm{SE}}aN2^{\frac{\eta_{\mathrm{SE}}}{N}}\ln 2}{\left(aN^2\left(2^{\frac{\eta_{\mathrm{SE}}}{N}} - 1\right) + bN + c\right)^2} = 0, \tag{5.7}$$

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

where $a = \frac{N_0}{L\eta_{PA}}, b = \frac{P_0}{W}$, and $c = \frac{P_{\text{common}} + L P_{\text{rf\_circuit}}}{W}$.

Combining (5.7) with (5.6), the relationship between the maximum EE η_{EE}^* and corresponding SE η_{SE}^* is

$$\eta_{\text{EE}}^* = \left(\frac{n_0 N 2^{\frac{\eta_{\text{SE}}^*}{N}} \ln 2}{L\eta_{\text{PA}}} \right)^{-1}. \tag{5.8}$$

The relationship between η_{EE}^* and η_{SE}^* is further given as

$$\lg\left(\eta_{\text{EE}}^*\right) = -\frac{\lg 2}{N}\eta_{\text{SE}}^* + \lg\left(\frac{L\eta_{\text{PA}}}{n_0 N \ln 2}\right), \tag{5.9}$$

which indicates that $\log \eta_{EE}^*$ scales linearly with η_{SE}^* and has a slope of $-\log 2/N$. Similar to the EE–SE relationship in classic Shannon theory [53], higher η_{SE}^* always leads to lower η_{EE}^*. The relationship between η_{EE}^* and η_{SE}^* is independent of P_0, P_{common}, $P_{\text{rf\_circuit}}$, and W, though as can be seen from (5.6), η_{SE}^* and η_{EE}^* are determined on the basis of all the other parameters.

In the case of independent N and M, the relationship is

$$\eta_{EE}^* = \left(\frac{n_0}{\eta_{PA} M} 2^{\frac{\eta_{SE}^*}{N}} \ln 2 \right)^{-1}. \tag{5.10}$$

For each case, there exists only one η_{SE}^* where EE monotonically increases with SE when SE is smaller than η_{SE}^*, and monotonically decreases with SE when SE is larger than η_{SE}^* [36].

It is expected, therefore, that the system operates at the green point. Also, it is important that η_{SE}^* satisfies the system SE requirement, and η_{EE}^* should be high enough. These require careful design of P_0, P_{common}, $P_{\text{rf\_circuit}}$, W, η_{PA}, N, and M. For example, when other parameters are given, M or N can be designed to maximize EE.

Optimal M for Maximizing EE for a Given SE, with Independent N and M

It is of practical importance to know how M affects EE for a given SE. If there is one optimal M that results in the highest EE, it is not necessary to implement too many antennas per transceiver. In the following exploration, we derive the optimal M to maximize system EE. Denote the denominator of (5.6) as $f(M)$:

$$f(M) = \left(2^{\frac{\eta_{\text{SE}}}{N}} - 1\right) \frac{N_0}{\eta_{\text{PA}}} \frac{N}{M} + \frac{N P_0 + P_{\text{common}} + NM P_{\text{rf\_circuit}}}{W} \tag{5.11}$$

The first- and second-order derivatives of $f(M)$ are

$$f'(M) = \frac{N P_{\text{rf\_circuit}}}{W} - \left(2^{\frac{\eta_{SE}}{N}} - 1\right) \frac{N_0}{\eta_{\text{PA}}} \frac{N}{M^2} \tag{5.12}$$

and

$$f''(M) = 2\left(2^{\frac{\eta_{SE}}{N}} - 1\right) \frac{N_0}{\eta_{\text{PA}}} \frac{N}{M^3} \geq 0 \tag{5.13}$$

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

Then $f(M)$ is a quasi-convex function of M. The optimal M^* that gives the minimum $f(M)$ is derived by making $f'(M) = 0$:

$$M^* = \sqrt{\frac{WN_0}{\eta_{\text{PA}} P_{\text{rf\_circuit}}} \left(2^{\frac{\eta_{\text{SE}}}{N}} - 1\right)} \tag{5.14}$$

Because of the definition of η_{EE} in (5.6), EE is a quasi-concave function of M, and the EE is at a maximum when $M = M^*$. When $M < M^*$, EE monotonically increases with M. When $M > M^*$, EE monotonically decreases with M. In practical system design, for a given SE there is one optimal number of antennas per transceiver that results in the highest EE. As in (5.14), the optimal M^* increases with SE and bandwidth, but decreases with PA power efficiency and $P_{\text{rf\_circuit}}$. For a given number of tranceivers N, more antennas per transceiver are needed for higher SE. If W increases, the noise power increases correspondingly, and a larger M is needed to achieve the SE. A larger $P_{\text{rf\_circuit}}$, however, reduces the optimal M^* because the increased circuit power may reduce EE.

Optimal N for Maximizing EE for a Given SE, with Independent N and M

In order to achieve the performance promised by massive MIMO, a large enough M is required. But the practical implementation of an equal number of transceivers is not trivial with many unresolved issues, including calibration and complexity. When the required SE is predetermined, it's very important to know whether a larger N always brings a better EE. Again, take the denominator of (5.6) as $f(N)$, the first-order derivative of $f(N)$ over N is derived as

$$\begin{aligned} f'(N) &= \frac{MP_{rf\_circuit} + P_0}{W} + \frac{N_0}{\eta_{PA}} \frac{1}{M} \left(\left(2^{\frac{\eta_{SE}}{N}} - 1\right) - 2^{\frac{\eta_{SE}}{N}} \frac{\eta_{SE}}{N} \ln 2 \right) \\ &= g\left(\frac{\eta_{SE}}{N}\right) = g(x). \end{aligned} \tag{5.15}$$

We have

$$g'(x) = -\frac{N_0}{\eta_{PA}} \frac{1}{M} \left(x 2^x (\ln 2)^2\right) < 0. \tag{5.16}$$

We also find that $g(0) = \left(MP_{rf\_circuit} + P_0\right)/W$ and $g(\infty) = -\infty$. This indicates that there exists only one x_0, such that $g(x_0) = 0$. Correspondingly, there exists only one N_0, $N_0 = \eta_{SE}/x_0$. When $N < N_0$, EE is monotonically increasing with N, and when $N \geq N_0$, EE is monotonically decreasing with N.

5.4.4 Standardization

The mmWave band communication technologies have been standardized by multiple international organizations. For example, the IEEE 802.11ad amendment to the 802.11 standard defines a directional communication scheme that takes advantage of beamforming antenna gain to cope with increased attenuation in the 60 GHz band. Beamforming training is used to determine the appropriate receive and transmit antenna sectors for a pairing of BS and UE. The training procedure is split into two subphases. During the

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

cell-specific beam sweeping, an initial coarse beam (or antenna sector configuration) is determined, which is used in a subsequent optional beam refinement phase. This is a simplified version of hybrid beamforming, in which the analog beamformer of the hybrid beamforming is fixed. Another example is IEEE 802.15.3c, which specifies the physical layer and MAC layer protocols and procedures for indoor wireless personal area networks (WPANs). As multiple antennas are available at both the transmitter and the receiver, codebook-based MIMO beamforming is employed.

The most recent standardization activity of hybrid beamforming technique is in 3GPP 5G NR. It has been agreed that this technology will be deployed in future 5G systems in the conference of 3GPP RAN1 #85 [54]. In this conference, the maximal number of the RF chains for the 5G BS and UE was determined as 32 and 8, and the maximal number of the antenna elements for the 5G BS and UE was 1024 at 70 GHz (128 @ 4 GHz and 256 @ 30 GHz) and 64 at 70 GHz (8 @ 4 GHz and 32 @ 30 GHz), respectively. Besides, the BS hybrid beamforming architecture is likely to employ the Architecture 2-A and the hybrid beamforming architecture of UE tends to be the Architecture 2-B in Fig. 5.11. In the conference of 3GPP RAN1 #86 [55], the DL beam management for hybrid beamforming was agreed upon. Specifically, the UE makes measurements on different BS transmit beams to support beam selection of BS and UE. In the conference discussion of 3GPP RAN1 #86b [56], the key problem of hybrid beamforming was to determine the main process of the beamforming procedure.

In order to guarantee the success probability of data transmission, it is necessary to choose proper TRP/UE beams, which means both analog beamforming and digital beamforming should be implemented properly for the BS and UE. The detailed DL CSI

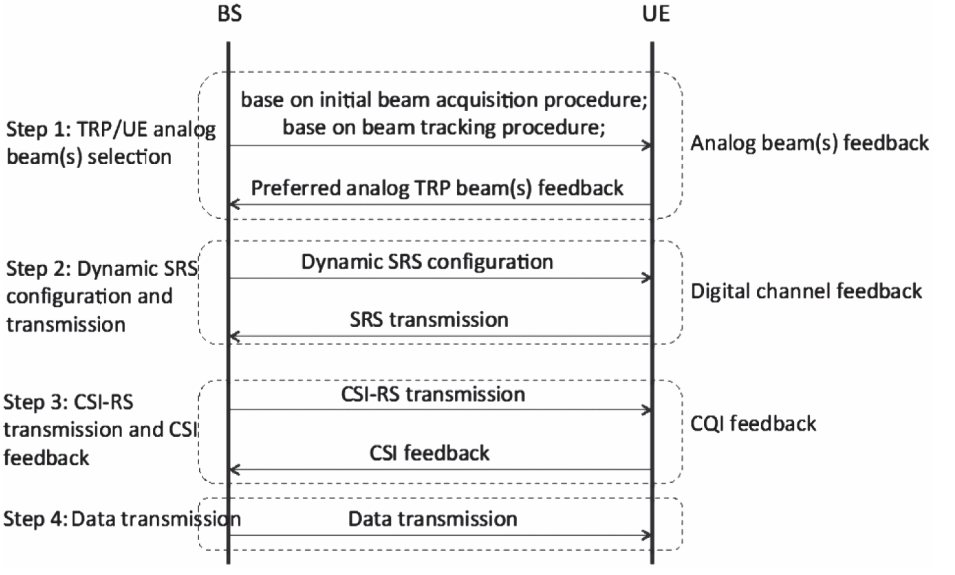


Figure 5.12 DL CSI acquisition framework for NR massive MIMO.

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

acquisition framework of NR MIMO, which is illustrated in Fig. 5.12, could be divided into the following 4 steps:

1. **TRP/UE analog beam(s) selection**
   First of all, the TRP/UE analog beam(s) should be selected properly to provide appropriate analog beamforming gain. Concretely, analog beam(s) could be selected based on initial beam acquisition procedure and/or beam tracking procedure. Here, we discuss these two analog beam selection schemes:

   - Scheme 1: Analog beam(s) are selected based on an initial beam acquisition procedure. For example, the TRP transmits synchronization signals and/or system information with the beam sweeping method, and a terminal would calculate and compare the power of different beams to identify its preferred TRP beam(s). Then the preferred TRP beam(s) information could be indicated to the network during or after the random access procedure explicitly or implicitly.
   - Scheme 2: Analog beam(s) are selected based on a beam tracking procedure. A beam tracking procedure is needed, since the preferred TRP/UE beam(s) may change when the channel condition between the TRP and UE changes. In order to facilitate the beam tracking procedure, a kind of beam selection RS needs to be designed. The beams carried by beam selection RS may be the same as or different from that carried by synchronization signals in the initial beam acquisition procedure. From the perspective of specification, these two kinds of beams, which applied in the initial beam acquisition procedure and beam tracking procedure separately, can be independent.

   It should be noted that the UE may need to feed back more than one preferred TRP beams in order to support SU-MIMO and MU-MIMO flexibly.
2. **Dynamic SRS configuration and transmission**
   After the base station receives the information of a UE's preferred analog beam(s), it could dynamically configure the UE to transmit UL SRS on some specific time/frequency/code resources. Then the base station will adjust its analog beam on these time/frequency/code resources to receive the UE's SRS. The applied TRP analog beams on these resources are transparent to the UE, and they could be flexibly adjusted not only according to the preferred analog beams fed back in step 1 but also according to the MIMO schemes (e.g., SU-MIMO or MU-MIMO) and/or multi-user pairing schemes. Both periodic and aperiodic SRS transmission could be further investigated.
3. **CSI-RS transmission and CSI feedback**
   For TDD systems, the BS can derive the DL channel information after step 2, and then the proper digital beamforming and rank information can be calculated. However, in order to determine the proper MCS, the BS needs to get the UE's interference information, or CQI report. One method is that the BS can apply the beamforming matrix derived from the SRS to UE-specific beamformed CSI-RS,

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

and then the UE measures the beamformed CSI-RS to derive the CQI and/or PMI information.

4. **Data transmission**
After the analog TRP beam(s), digital beamforming matrix, rank information, and MCS are determined, the BS can transmit traffic data to the UE.

Note that this CSI acquisition framework for NR massive MIMO applies to all the scenarios, no matter what beamforming structure and RS pattern is utilized.

5.4.5 Summary

The diverse use cases and scenarios of 5G motivate software-defined and flexibly configured air interface. It has been a consensus that MIMO technique will be the cornerstone of 5G air interface. However, different beamforming structures may possibly be considered in the system deployment of 5G, including digital beamforming, analog beamforming, and hybrid beamforming, with various mappings between TXRU and antenna elements. This necessitates the design of flexible MIMO with a unified beamforming architecture and a unified CSI acquisition/feedback mechanism. To this end, this section has so far proposed a flexible MIMO framework for 5G communications. Also, various hybrid beamforming algorithms were surveyed for both full-array and sub-array hybrid beamforming architectures. Furthermore, the energy-efficient design of a sub-array structure was examined, with the optimal number of TXRU and antenna elements analyzed. Finally, the standardization of 3GPP 5G NR on hybrid beamforming was investigated, and a unified CSI acquisition and feedback framework was discussed, which is applicable to various beamforming structures and reference signal designs.

5.5 New Waveform

Driven ultimately by diversified applications, new waveform, as one key enabler of SDAI, is envisioned to be able to support various extreme requirements in the physical layer. Orthogonal frequency division multiplexing (OFDM) has been recognized as an effective waveform for mobile communications, due to its ease of implementation, robustness to multipath fading, and MIMO friendliness. Some shortcomings of the current solution emerge when coming across future requirements, including: not-very-well localized in time and frequency domain, and sensitivity to frequency or timing synchronization error, etc. For new radio (NR), several new waveform schemes attract industry's interests, including:

- Filter bank multi-carrier (FBMC),
- Universal filter multi-carrier (UFMC),
- Generalized frequency division multiplexing (GFDM),
- Filtered-OFDM (f-OFDM) and windowed OFDM (w-OFDM),
- Orthogonal Time Frequency Space (OTFS),
- Variants of DFT-S-OFDM, etc.

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

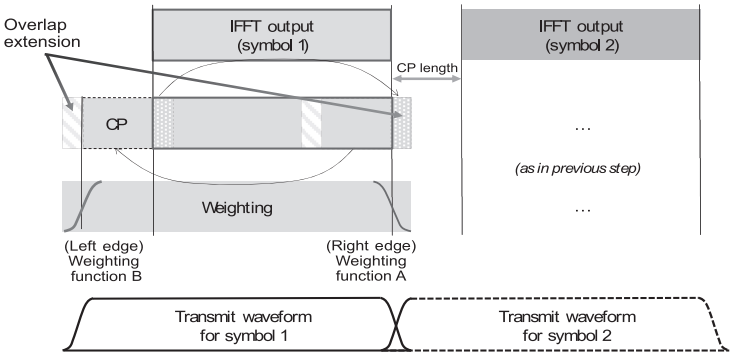


Figure 5.13 WOLA at transmitter with CP-OFDM.

In the following sections, we will briefly introduce the technical principles of each waveform scheme and provide the technical comparison of different solutions. Moreover, the progress of 3GPP 5G NR specification on waveform schemes is discussed, including why two waveforms, DFT-S-OFDM and OFDM, could be flexibly configured as a 5G NR UL waveform scheme (this never happened in previous telecom systems). A unified framework of waveform design is proposed, where multiple waveforms can be flexibly implemented according to the specific use scenario and channel conditions. This is particularly important to achieve green and with respect to green and soft 5G network operations.

5.5.1 w-OFDM/f-OFDM

To offset the poor frequency localization for CP-OFDM, an efficient spectrum-shaping technique of windowing or filtering approach can be utilized. W-OFDM is synthesized by a conventional CP-OFDM waveform, followed by a weighting and overlap-and-add (WOLA) operation. The better-contained frequency response is achieved by adding soft edges to the cyclic extension of the OFDM symbol in the time domain, as shown in Fig. 5.13. When the edges further expand, the overhead is still the same as a CP-OFDM waveform, since adjacent symbols are overlapped in the edge transition region. The shape of the window (or edge) in time domain determines the frequency response of the prototype filter. In general, a raised-cosine edge seems to offer a good compromise with straightforward implementation [57].

In principle, f-OFDM applies a filter with a subband of the CP-OFDM system to reduce the OOB leakage [58]. One example of filter design is the windowed sinc function-based method. To be specific, the transmit filter is computed as the product of the ideal band-pass filter and a time domain mask [59]:

$$f(n) = p_i(n)w(n),$$

where $p_i(n)$ is the ideal band-pass filter covering the allocated bandwidth of the i-th user, and $w(n)$ is a raised-cosine window with duration T_w. The window has smooth

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

transitions to zero on its both ends so that it avoids abrupt jumps at the beginning and end of the truncated filter. Furthermore, up to half symbol length can be used for T_w. The long filter length of $f(n)$ provides good OOB emission suppression. Different from WOLA, the band-pass filter of f-OFDM is bandwidth dependent. Therefore, the filters need to be constructed based on the tone allocation. Another concern of applying f-OFDM, especially for the TDD band, is the long group delay due to the long filter length.

5.5.2 UFMC

Similar to f-OFDM, UFMC is another spectrum-shaping technique utilizing the filtering approach. The main difference is in how the band-pass filter is constructed [60]. Specifically, a band-pass filter is carefully designed for a fixed bandwidth, e.g., a resource block (RB). The same filter can be universally reused only by shifting the center frequency. That is, when n RBs are assigned to the transmitter, n parallel IFFT and filtering operations have to be computed. Instead of CP, a guard interval (GI) filled with zeros is introduced between the symbols to prevent ISI due to filter delay. The filter length is set to be the same as the GI duration (usually not long as the filter in f-OFDM). Since GI is introduced instead of CP, the cyclic convolution property is not preserved in UFMC. Therefore, the receiver structure is not as simple as the one in CP-OFDM. Specifically, doubled-sized FFT is used at the receiver, but only the even tones of the doubled-sized FFT outputs are used for the detection, which increases complexity and latency of the decoder. The detailed operation of the transmitter and receiver could be found in Fig. 5.14.

5.5.3 FBMC

FBMC has drawn much interest due to its excellent spectral containment [61]. It is achieved by optimizing the shape of the prototype filter $p(n)$ through oversampled coefficients on the granularity of each carrier. The modulator and demodulator are conceptually illustrated in Fig. 5.15.

In FBMC, a prototype filter satisfying generalized Nyquist constraints is used for both signal synthesis at the transmitter and signal analysis at the receiver. Figure 5.16 shows a prototype filter with the oversampling factor $K = 4$, where K is denoted as

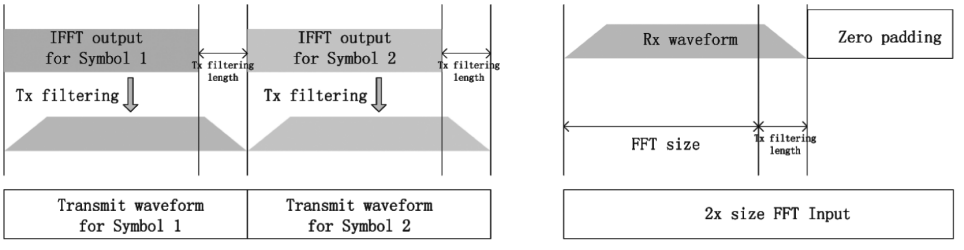


Figure 5.14 UFMC processing at the transmitter and receiver.

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

Table 5.2 Frequency domain coefficients of prototype filter.

H_{-3}	H_{-2}	H_{-1}	H_0	H_1	H_2	H_3
0.235147	$\sqrt{2}/2$	0.97196	1	0.97196	$\sqrt{2}/2$	0.235147

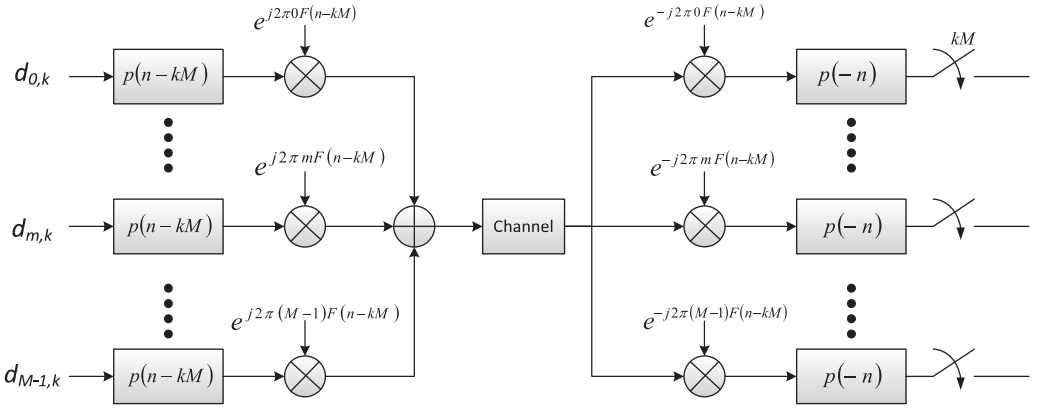


Figure 5.15 Modulator/demodulator of filter bank multi-carrier.

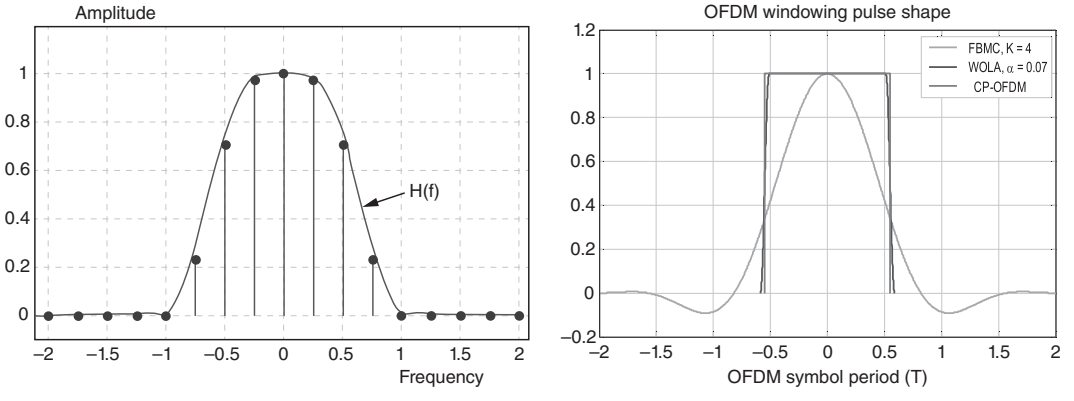


Figure 5.16 Illustration of the prototype filter (left/right: frequency/time domain response).

the FBMC overlapping factor. And the interval between adjacent coefficients is $1/4\triangle_f$, where $\triangle_f = 1/T$ is the sub-channel spacing. Further, because of the oversampled frequency coefficients, the prototype filter spans multiple symbol periods T, as shown in the right figure in Fig. 5.15.

The specified nonzero coefficients are summarized in Table 5.1. It can be verified that the selected frequency coefficients satisfy the Nyquist property.

Note that there are several limitations for practical implementation apart from the complexity issue. Special modulations such as OQAM may be necessary to avoid inter-channel interference introduced by the prototype. In the case of multipath channels,

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

the orthogonality statement at demodulator is no longer valid, since there is no CP protection in this waveform, and channel convolution is not precisely cyclic. Another potential limitation in applying FBMC is the deployment with MIMO when exploiting more degrees of freedom. In addition, reference signals may be less flexible than those in OFDM and hard to enable efficient channel estimation techniques [62].

5.5.4 GFDM

In GFDM, the prototype filter for each sub carrier is also specifically chosen to be well-localized in frequency domain to reduce out-of-band emission, similar as in FBMC. The main difference from FBMC is that, in GFDM: 1) multiple OFDM symbols are grouped into a block, with a CP added to the block; 2) within a block, the prototype filter is "cyclic-shift" in time, for different OFDM symbols [63]. A block of GFDM waveform can be expressed as:

$$x(n) = \sum_{k=0}^{K-1} \sum_{m=0}^{M-1} p_{k,m}(n)\, d_{k,m} \quad for\, n = 0, 1, \cdots, N-1. \tag{5.17}$$

Each block has $N = KM$ samples, which can be decomposed into M sub-symbols. Each M sub-symbol contains K subcarriers. The pulse $p_(k,m)(n)$ is the frequency and time-shifted version of the prototype filter $p(n)$, as shown in (5.18). Specifically, the module operation makes $p(n)$ circularly shifted in time by m sub-symbols, and the exponential term shifts the filter in frequency by k subcarriers.

$$p_{k,m}(n) = p\left[(n - mK)\, mod N\right] e^{j2\pi k \frac{n}{K}} \tag{5.18}$$

Figure 5.17 shows an example of GFDM resource partitioning with M sub-symbols per block, with time offset T/M between adjacent sub-symbols. The duration of each symbol can be longer than T/M due to the specially engineered prototype filter. Each sub-symbol contains $K = BT/M$ sub-channels, with spacing of M/T (Hz) spacing between adjacent sub-channels.

In order to avoid interference between sub-symbols within a block, the selected prototype filter should have Nyquist property. Further, special modulations such as OQAM may be necessary to avoid interchannel interference as in FBMC. Otherwise, complicated receiver algorithm is needed to handle the interference [64]. In addition, due to the cyclic structure of the block and the use of CP, GFDM improves the capability against with ISI with the penalty of spectral containment property especially when a waveform spanning multiple blocks in time.

5.5.5 OTFS

OTFS, instead of using filtering or windowing technologies for spectral containment, as above, characterizes the Doppler-induced time varying nature of the wireless channel and parameterizes it as a 2D impulse response in the delay-Doppler domain [65]. Specifically, the channel time-frequency response $H[n,m]$ is related to $h(\tau, v)$ via the

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

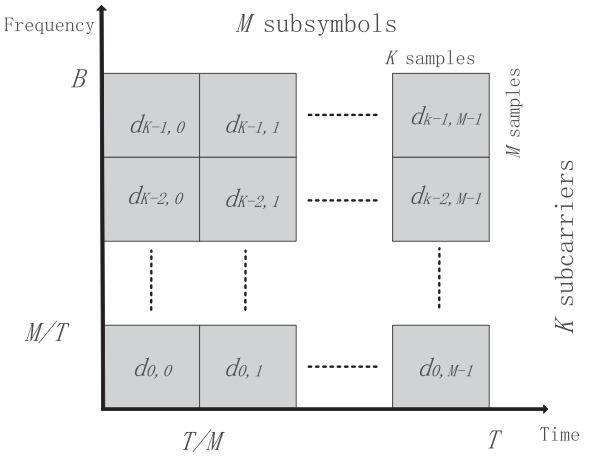


Figure 5.17 Resource partition in GFDM.

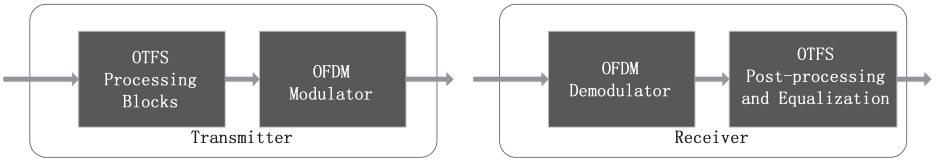


Figure 5.18 OTFS architecture with pre- and post-processing blocks.

transform shown in (5.19), where the delay-Doppler response $h(\tau,)$, but not time varying impulse response $h(\tau, t)$, is used for characterization of the channel.

$$H[n,m] = \iint^{h} (\tau, \nu)\, e^{j2\pi\nu nT} e^{-j2\pi m\Delta f\tau} d\nu d\tau, \tag{5.19}$$

where τ and ν denote delay and Doppler respectively, m and n are for the frequency bins and time bins, T is the length of the OFDM symbol (plus CP extension). Above equation can be thought of as a 2D Fourier transform version of the delay-Doppler impulse response $h(\tau,)$. More details on the theory analysis of OTFS can be found in [66].

In terms of implementation, 2D OTFS consists of a DFT along the delay/frequency dimension and an IDFT along the Doppler/time dimension. The transformation consists of pre- and post-processing blocks in the transmitter and receiver respectively, as depicted in Fig. 5.6. This block diagram is analogous to the blocks used to implement DFT-s-OFDM on top of an underlying OFDM signal chain. The pre- and post-processing blocks could enable QAM modulation in the delay-Doppler domain. In this way, all QAM symbols experience the full diversity of the channel. Further, at a high-mobility scenario, the time invariance property holds for the duration of the TTI. It would make a closed-form transmission mode possible for spectral efficiency (SE) improvement.

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

No doubt that additional preprocessing blocks introduce higher complexity at both the transmitter and receiver. Another concern is the longer processing latency when the preprocessing block spans a TTI length or multiple OFDM symbol duration. The receiver has to wait until getting the last symbol before it can go on its next step operation, which may be intolerable for the real system.

5.5.6 Variants of DFT-s-OFDM

DFT-s-OFDM, featured by a high-power efficiency due to its low PAPR, is very well known as the UL waveform scheme of LTE. Further, some variants of DFT-s-OFDM, including unique word (UW) DFT-S-OFDM, are designed for superior OOB suppression performance [67].

The transmitter structure for UW-DFT-s-OFDM is illustrated as Fig. 5.19, where a unique word is added prior to the DFT operation. It leads to suppression of out-of-band emission, thanks to the cyclic property obtained from DFT and IDFT operation. Also, unique word at the output of IDFT serves as a guard between two data parts of consecutive OFDM symbols, thus, CP insertion is not necessary. In addition, it allows the possibility of UW-based time domain channel estimation and synchronization design. Note that insertion of UW does not change PAPR of DFT-s-OFDM, as long as the envelope of UW is nearly constant or similar to that of data symbols. Further, if UW is

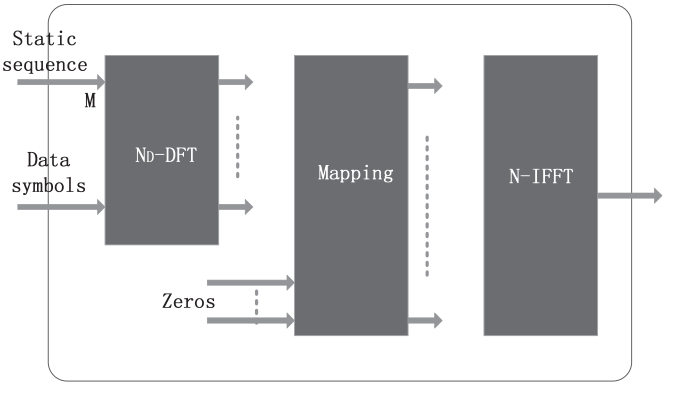


Figure 5.19 Transmitter diagram for UW-DFT-s-OFDM.

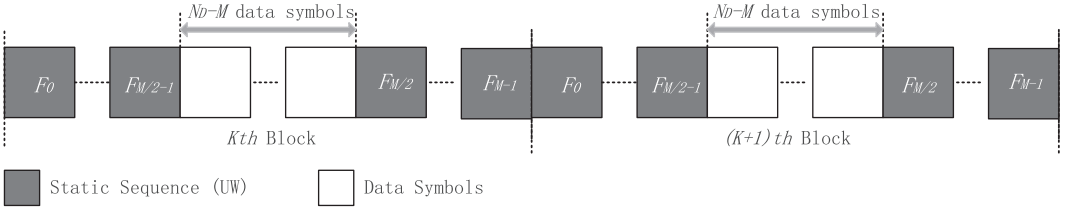


Figure 5.20 Placement of UW in a block.

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

set to be a zero sequence, i.e., zero-tail (ZT) DFT-S-OFDM, better spectral confinement is expected at the price of PAPR [68].

An example of UW sequence is placed as shown in Fig. 5.20. The number of symbols in UW is denoted by M, and UW is split in half and placed at the head and tail of the DFT-s-OFDM block.

5.5.7 Constant Envelope Waveform

A simple way to achieve high transmit efficiency is to employ a constant envelope waveform, which allows almost any PA to operate at saturation point. Minimum shift-keying (MSK) and Gaussian MSK (GMSK) are the most popular constant envelope waveforms. MSK can equivalently be viewed as offset-QPSK (quadrature phase shift-keying) with sinusoid pulse shaping, which provides efficient modulation and demodulation [69]. Notice that a differential encoder is inserted before the modulator to avoid error propagation at the demodulator. GMSK is a variant of MSK, where a Gaussian-filtered version of the information sequence is applied to an MSK modulator [70]. The Gaussian filter helps to increase the SE of the MSK, with reduced inter-symbol interference. Note that with the introduction of Gaussian filtering, the GMSK signal can no longer be viewed as offset-QPSK. The drawback, however, is the inefficiency from a capacity perspective compared to QAM. But for applications like low data-rate packet transmission in IoT, a constant envelope waveform may be attractive, since it achieves the highest PA efficiency.

5.5.8 Unified Waveform Framework

A common feature of the above new waveforms is that filters are employed to suppress the out of band emission and relax the requirements on time-frequency synchronization. But there are also subtle differences among these waveforms. The filters in UFMC and f-OFDM are implemented at the granularity of each sub-band. The main difference is that f-OFDM uses a longer filter and the signal processing procedure is same as the conventional OFDM in each sub-band for backward compatibility. In contrast, UFMC uses a shorter filter, and the CP of OFDM is replaced with an empty guard period. GFDM can cover CP-OFDM and SC-FDE, which can be regarded as special cases, according to different numbers of subcarriers and sub-symbols in a GFDM block. In addition, the overhead is kept small by adding CP for an entire block that contains multiple sub-symbols. The filter in FBMC is implemented at the granularity of each subcarrier. By a well-designed prototype filter, FBMC can greatly suppress side-lobes of a signal. Moreover, the overhead can also be reduced by removing CP in FBMC as UFMC. At the same time, in order to reduce the interference of adjacent sub-channels and computation complexity, the OQAM modulation and polyphase network is needed in FBMC and GFDM schemes. In addition to the above waveforms, DFT-S-OFDM is also supported by the 3GPP 5G NR as one of the UL waveforms for the improvement of PAPR and UL coverage improvement. In order to better embrace the challenges of fast fading in high-mobility scenarios, e.g., high speed train, OTFS is proposed, which

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

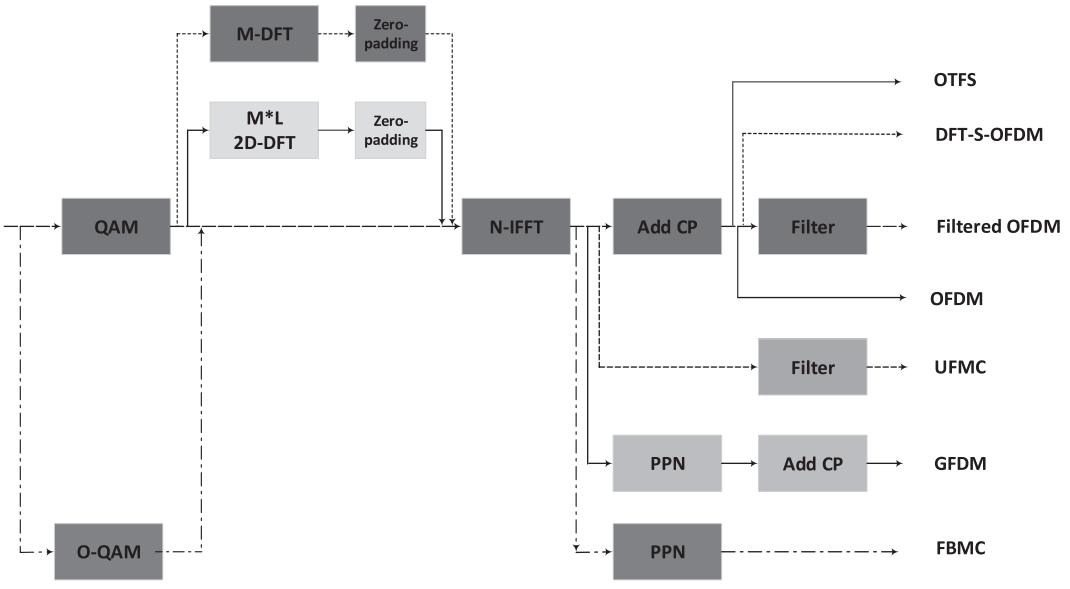


Figure 5.21 A unified framework of waveforms.

transforms the time-varying multipath channel into a time invariant delay-Doppler channel. The transmitter structure of the OTFS is similar to that of the DFT-S-OFDM, except for the 2D-DFT processing.

A unified framework to implement various waveforms is shown in Fig. 5.21, where the waveform can be represented as (5.20),

$$x(t) = \sum_{u \in U} \sum_{k \in K_u} \sum_{n=-\infty}^{+\infty} \sum_{m=1}^{M} s_{k,n}(m) g_{k,m}(t - nT) e^{j2\pi f_k (t-nT)} \otimes h_u(t) \tag{5.20}$$

where $s_{k,n}(m)$ is the m-th sub-symbol in the n-th transmission symbol and k-th subcarrier. $g_{k,m}(t)$ is the shaping filter in a single symbol. The filter of each user is denoted as $h_u(t)$. And the frequency of subcarrier is denoted as f_k. The symbol duration is T. And the convolution operator is denoted as $\otimes$.

By the unified structure, we can flexibly configure different waveform schemes according to various 5G scenarios on the basis of minimizing the hardware functional module. For instance, if $g_{k,m}(t)$ is a rectangular window and with length T, $M = 1$, $h_u(t) = \delta(t)$, $f_k = k/T_s$ and T_s is the symbol length excluding CP, $x(t)$ is actually the OFDM signal. When $M \neq 1$, $g_{k,m}(t) = g[(t - mK) \bmod (KM)]$, $g(t)$ is the prototype filter and the other parameters set same as the above OFDM parameters, $x(t)$ becomes a GFDM signal.

Based on the above waveform framework, multiple waveforms can be flexibly implemented accordingly to the specific use scenario and channel condition. Note that

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

some waveforms can be implemented without standardization efforts, e.g., f-OFDM, w-OFDM, and UFMC.

5.5.9 Waveform for 5G NR in 3GPP

DL Waveform for NR

For NR DL, OFDM-based waveforms are preferred as the candidate schemes. Specifically, OFDM is identified as the NR DL waveform, while filtering or windowing approach, i.e., f-OFDM or w-OFDM, is used for efficient spectrum-shaping to reduce the in-band and out-of-band emission. It aims to reach 98% spectrum efficiency of the bandwidth in NR while LTE reserves up to 10% of the bandwidth as guard bands to abide by the spectrum mask [71]. It means that a particular filtering or windowing function may be needed, as discussed in the previous subsection. Note that it is an implementation issue for this specific method, from the perspective of 3GPP RAN1's perspective. In the following, the major performance indicators are listed for NR DL waveform [72]. These factors determine the selection of waveform scheme.

- **Spectral efficiency**: To meet extreme data rate requirements for both DL and UL. In general, SE is more important at lower carrier frequencies than at higher frequencies, since the spectrum is not as precious at higher frequencies due to the availability of potentially much larger channel bandwidths.
- **MIMO compatibility**: To enable a straightforward use of MIMO technology. Multiple or massive antenna technology is considered as one of key enablers for data-rate boosting (up to 20Gbps for peak data rate) and coverage improving. Thus, the new waveform should have limited implementation complexity with MIMO integration.
- **Transceiver baseband complexity**: To enable efficient baseband processing at large bandwidths envisioned for NR. Reasonable implementation complexity should be involved for not only the waveform itself, but also its related implementation method too. Signal detection and channel estimation/equalization at the receiver should not have very high complexity. Note that at very high frequencies, the receiver may also have to cope with severe RF impairments.
- **Flexible numerology configuration**: To enable different services (with different numerologies) simultaneously on the same carrier. For example, the uRLLC or synchronization signal may require larger subcarrier spacing for shorter transmission time interval.
- **Frequency localization**: To support the coexistence of different services that are potentially enabled by mixing different numerologies in frequency domain on the same carrier. Further, it is essential to provide minimal loss in SE. Also, efficient asynchronous communication would require a waveform with minimal inter-UE interference leakage, which is achieved by good frequency localization.
- **Time localization**: To efficiently enable (dynamic) TDD and support latency critical applications such as uRLLC. Low latency is very important for all link types.

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

- **Robustness to synchronization errors**: This important where synchronization is hard to achieve, such as a D2D link.
- **Robustness to channel time-selectivity**: This is important in high speed scenarios. Fast time-varying characteristics would make channel tracking difficult. More reference signals (higher overhead) or additional function blocks (higher complexity) may be needed for robust channel estimation.
- **Robustness to channel frequency-selectivity**: This is always an important measure in multipath channels. Channel frequency selectivity depends on various factors as type of deployment, beamforming technique, and bandwidth. It is important that BSs can cope with frequency selective channels without complicated receiver limitations.
- **Robustness to phase noise**: This is important for all link types especially for a high-frequency device (transmitter/receiver), as phase noise typically increases with carrier frequency. Note that high-quality oscillators mean high cost and may not be affordable.
- **Low cubic metric**: To compensate for the PA's inefficiency. A low cubic metric (or PAPR) is important for power efficient transmissions, and becomes even more important at very high frequencies. Note that small-sized, low-cost BSs are envisioned at high frequencies, therefore, a low cubic metric is also important for DL.
- **Flexibility/scalability**: To support diverse services in wide range of frequencies.

Taking the above factors into account, OFDM shows overwhelming advantages. It reaches the consensus without much debates in 3GPP that OFDM (with filtering or windowing) is identified as a NR DL waveform. Note that for mMTC, high frequency or high mobility, it still retains some possibility for other solutions.

UL Waveform for NR

OFDM-based waveforms are also considered as the candidates for NR UL. Different from NR DL or other communication systems, two schemes, OFDM and DFT-S-OFDM, are both identified as the UL waveform. Specifically, DFT-S-OFDM and OFDM could be flexibly configured by the network. For cell-edge users, DFT-S-OFDM can be configured to improve the coverage. While for users with good channel conditions and multiple antenna ports, OFDM could be used by combining with multi-layered transmissions for data rate boosting. In the following arguments, a general comparison of two waveforms for UL are presented first, and then more detailed technological principles are provided in the next subsections.

Why DFT-S-OFDM is selected for NR UL waveform

1. **PAPR/cubic metric**

   Besides PAPR, a cubic metric is a more accurate measurement metric for the power back-off required for amplifiers [73, 74]. Table 5.3 summarizes the PAPR/CM for OFDM, DFT-S-OFDM, and OFDM with PAPR reduction technology [73].

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

Table 5.3 PAPR/CM comparison of OFDM, DFT-S-OFDM and OFDM with companding.

	OFDM			DFT-S-OFDM			OFDM with PAPR reduction		
	QPSK	16 QAM	64QAM	QPSK	16 QAM	64QAM	QPSK	16 QAM	64QAM
PAPR(0.1%)	10.614	10.571	10.665	7.446	8.403	8.626	6.432	6.989	7.082
Cubic Metric	3.29	3.31	3.32	1.02	1.80	1.95	2.54	2.58	2.59

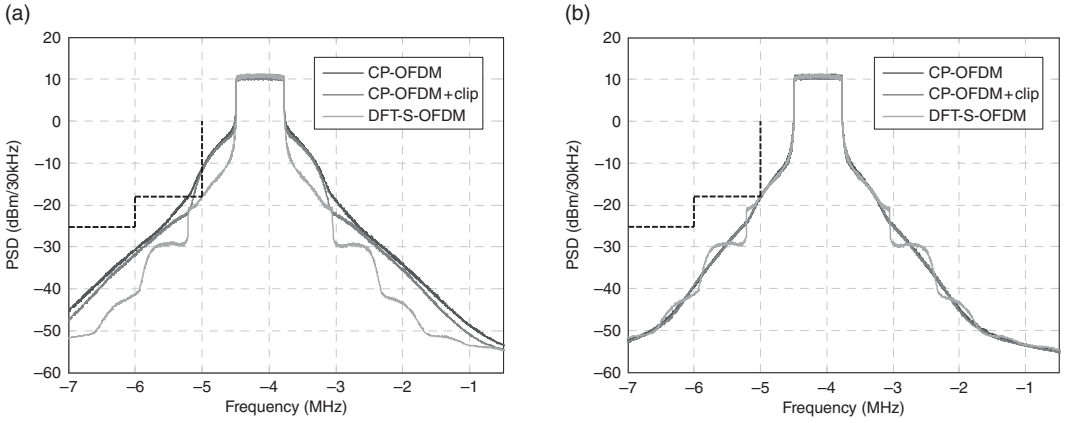


Figure 5.22 OOB emissions performance, where (a): without power back-off; (b): with power back-off.

Note: 4 RBs and subcarrier spacing at 15kHz is assumed.

Both DFT-S-OFDM and OFDM with PAPR reduction technology could give a low PAPR, while DFT-S-OFDM has the lowest cubic metric. That is, in theory, DFT-S-OFDM has the lowest power back-off value, i.e., the maximal output power.

2. **Power back-off**

Figure 5.22 presents the power spectrum density (PSD) for the just-discussed three waveforms without and with power back-off, respectively. The PA model in [75] is considered with post-PA loss as 4dB. Figure 5.22a shows that when using maximal output power [23dBm], PSD of OFDM is out of the emission mask (black dash). Then, by setting different values of power back-off in Table 5.4, all the waveforms can satisfy the requirements of in-band and out-of-band (OOB) emissions (including the reuse of the ACLR and UE emission mask in TS 36.101 for LTE [76]), illustrated in Fig. 5.22b.

Further, Table 5.5 shows the EVM performance, considering both with and without power back-off. With power back-off as described, the EVM fulfills the maximum tolerable limit for QPSK modulation as defined by 3GPP, i.e., 17.5% as in LTE systems [76].

Note: The power back-off values are set the same as Table 5.4.

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

Table 5.4 Power back-off for in-band and out-of-band emission requirements.

Schemes	Power back-off [dB]	Output power (post-PA loss = 4dB)
DFT-S-OFDM	0	23dBm
OFDM	−2.0	21dBm
OFDM with PAPR reduction	−1.5	21.5dBm

Table 5.5 EVM w/ or w/o power back-off.

Schemes	EVM (w/o power back-off)	EVM (w/ power back-off)
DFT-S-OFDM	7.36%	7.36%
OFDM	22.97%	9.97%
OFDM with PAPR reduction	23.95%	13.5%

Based on the agreed-upon PA model, DFT-S-OFDM could reach the maximal output power [23dBm], and OFDM's maximal output power is reduced to 21.5 dBm. Note that a 1.5dB power gap always exists no matter the transmission scheme or frequency band. Also, the 1.5dB link budget corresponds to a 40m to 50m distance [77] around 4GHz, which is actually a large coverage range. Besides, as no terminal/chip manufacturers would like to use new PAs with larger linear regions, DFT-S-OFDM is identified as the waveform scheme of NR UL.

Why OFDM is selected for NR UL waveform

The BLER vs. SNR curves for different waveforms are presented in Fig. 5.23. It shows that for QPSK, only 0.2dB gain for OFDM is shown. At higher MCS (16QAM/64QAM), OFDM outperforms DFT-S-OFDM greatly, with a gain of 1.5–2dB. As described, DFT-S-OFDM could provide a 1.5dB gain for maximal output power compared with OFDM. But for a cell-center user with high SNR, the gain of DFT-S-OFDM vanishes due to the BLER performance gap for these two waveforms. Furthermore, DFT-S-OFDM may suffer performance loss when no maximal output power is set. For example, if the UE output power is set to be 18dBm, OFDM provides nearly 2dB gain compared with DFT-S-OFDM.

Furthermore, OFDM is preferred due to its friendliness to MIMO. For dynamic TDD, symmetric DL/UL link allows for the possibility of crosslink interference mitigation. Therefore, OFDM is also selected as the waveform scheme of NR UL.

In summary, OFDM used in the UL (and also in side-links) comes with several advantages, as listed below:

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

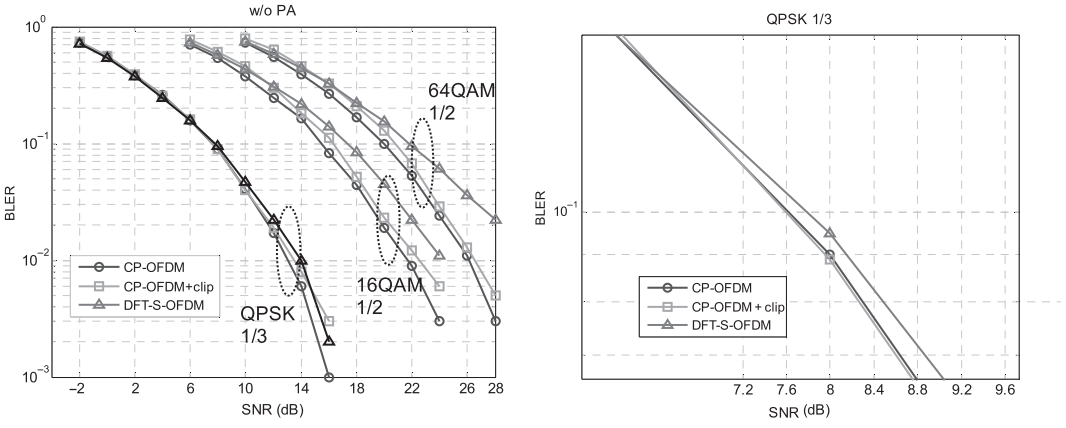


Figure 5.23 BLER performance for OFDM, DFT-S-OFDM, and OFDM with companding.

- It opens up for a more flexible UL scheduling. Having the same transmission scheme in both UL and DL makes the whole system design symmetrical. Further, it simplifies the overall system design by reducing the need for specific baseband receivers for respective link.
- It helps to facilitate the UL MIMO feature in NR. OFDM has shown significant advantages when considering multilayer transmission and hence, OFDM is preferred for UL MIMO use cases.

The agreement is that OFDM is also identified as the waveform scheme of NR UL. In a practical sense, the network can decide and communicate to the UE which CP-OFDM- and DFT-S-OFDM-based waveform to use. Further more, a common framework is targeted in designing CP-OFDM- and DFT-S-OFDM-based waveforms.

5.5.10 Summary

We focused on the discussion of new waveforms in 5G NR in this section. Several new waveform schemes were provided, with technical principles of each scheme and corresponding advantages and disadvantages. Then, the standardization progress of waveform schemes in 3GPP were discussed. For NR DL, OFDM with filtering or windowing was identified as a DL scheme. For NR UL, two waveforms, DFT-S-OFDM and OFDM, could be flexibly configured by the network. The technical principles behind the decision are presented. For mMTC/high frequency/high mobility, there is still some possibility for other solutions. Under the framework of SDAI and the framework of waveform design, various schemes, including those specified in the 5G NR standard and UE-transparent schemes, can be flexibly configured for different scenarios and applications.

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

5.6 Flexible Multiple Access Schemes

The current wireless communications systems have predominantly adopted orthogonal multiple access (OMA) schemes, where users are allocated with orthogonal radio resources in time, frequency, or space domain. Existing OMA schemes are able to efficiently eliminate multiuser interferences and thus allow relatively simple transceiver implementations. However, to the multiuser case, it is shown that OMA schemes achieve lower capacity than non-orthogonal schemes in the DL broadcast channel (BC) and the UL multiple access channel (MAC). Such inefficiency of OMA schemes is exacerbated in the UL scenario. Utilizing the channel based on existing OMA schemes may lead to a severe waste of radio resources, or even fail to work in massive connectivity scenarios, such as IoT applications.

The design of the 5G radio network is aiming for higher capacity, larger connectivity, and lower latency [78], which should provide better user experience for eMBB, mMTC, and URLLC services. The mMTC application scenario target to support a massive number of devices simultaneously while the URLLC scenario enables mission-critical transmissions with ultra-high reliability and ultra-low latency. Toward these goals, non-orthogonal multiple access (NoMA) opens the horizon for a new angle of thinking. As has been predicted by multiuser information theory, system capacity can be greatly improved by NoMA transmission compared with that of OMA transmission. So it is very suitable for the uplink massive number of simultaneous users. Besides, as a collision resolution method, NoMA combined with grant-free transmission can also improve the reliability and reduce the latency. And due to their non-orthogonal nature, the requirement of precise channel feedback and scheduling for multiuser multiplexing is thus reduced, or even removed in some scenarios. So, since NoMA-based, grant-free transmission schemes have potential advantages over Orthogonal MA in the aspect of collision resolution or robustness in a resource-limited scenario, these schemes may be solutions to future applications that have very stringent latency requirements, e.g., URLLC, etc.

In all, considering the above diverse requirements of different scenarios, the flexible OMA and NoMA schemes should be introduced in the future 5G NR system.

This section starts with an introduction of some typical NoMA schemes, which are under discussion in 3GPP 5G NR standardization, followed by some theoretical analysis of a NoMA system. Then a flexible MA structure is presented to meet the requirements of diversified services.

5.6.1 Potential New Multiple Access Techniques for 5G

NoMA Based on Super Position Coding

Superposition coding-based non-orthogonal multiple access (SPC-NoMA) utilizes a power domain for user multiplexing and can be applied for both DL and UL. Established by network information theory, non-orthogonal access with SIC/DPC can achieve the multiuser capacity region both in UL and DL. SPC-NoMA superposes multiple users in

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

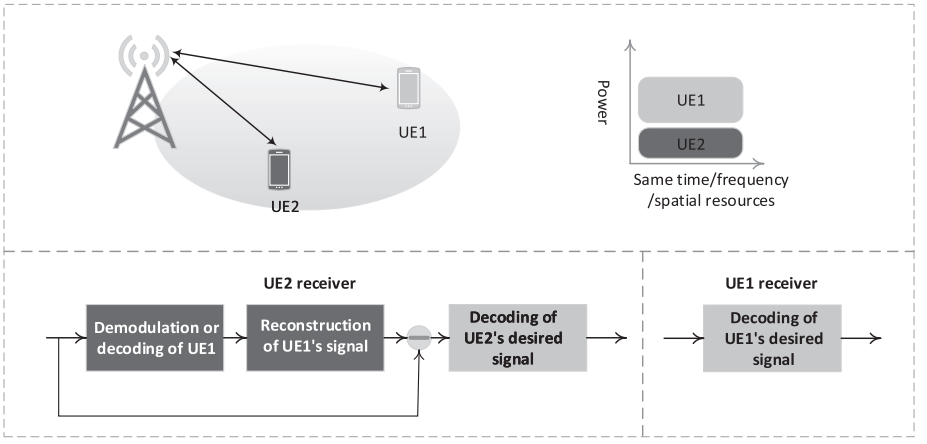


Figure 5.24 Illustration of SPC-based NOMA and transmitter/receiver.

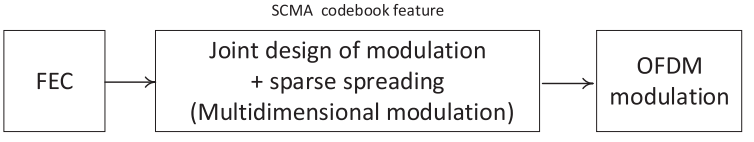


Figure 5.25 Abstracted SCMA transmit procedure for each data layer.

the power domain and exploits channel gain difference between the multiplexed users with the aid of an advanced receiver, e.g., a successive interference cancellation (SIC) receiver, for user separation. Figure 5.24 shows signal transmission and receiving in DL SPC-NoMA system with two users. In the SPC-NoMA system, at the transmitter side, the complex modulated symbols of different UEs are superposed with different transmission power settings. At the receiver side, the symbols of different UEs can be recovered by interference cancellation. With the joint optimization of transmitter and receiver in the SPC-NoMA, multiple layers of data can be simultaneously delivered in the same time, frequency, and spatial resource. SPC-NoMA techniques were discussed in 3GPP under the study item of "study on DL multiuser superposition transmission" in release 13, which was confined to DL data transmission. For the 5G system, more application scenarios of SPC-NoMA techniques, such as UL and control channel, and more advanced SPC-NoMA techniques, such as the combination with MIMO techniques and intercell mitigation schemes, are investigated.

SCMA

SCMA is a novel MA technique. It maps coded bits of a data stream to a sparse codeword of a codebook built based on a multidimensional constellation.

As shown in Fig. 5.25, at each SCMA layer, the SCMA modulator maps input bits to a complex multidimensional codeword selected from a layer-specific SCMA codebook, which has its own sparsity pattern (location of nonzero entries). One or multiple SCMA layers can be assigned to a user/data stream.

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

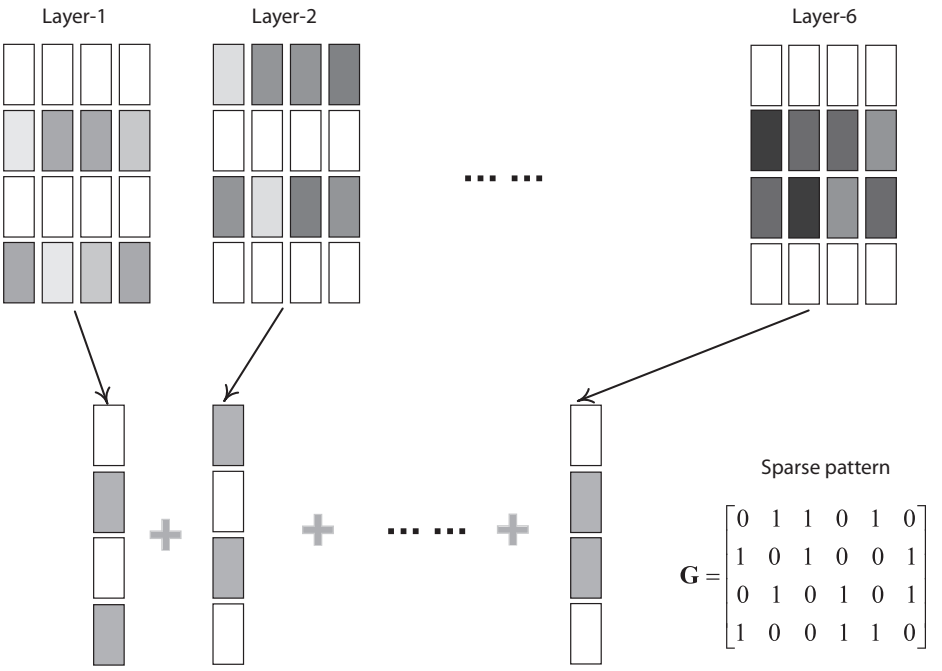


Figure 5.26 SCMA codebook illustration: bit-to-codeword mapping.

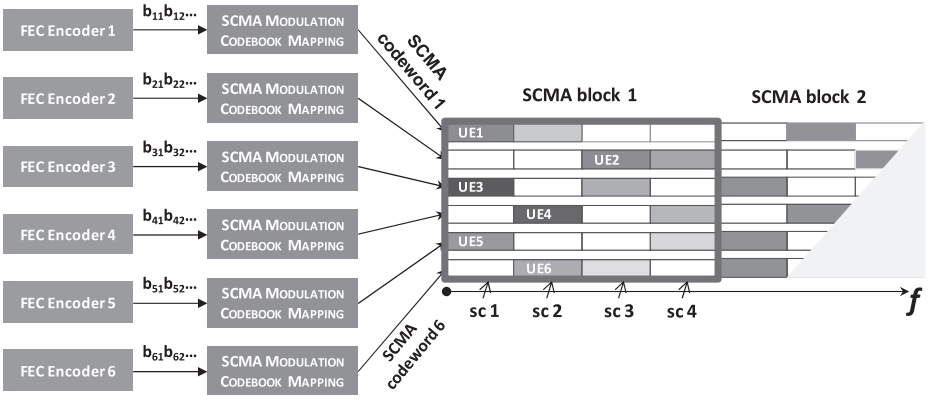


Figure 5.27 Illustrative features for SCMA.

Figure 5.26 shows an example of a codebook set with six data layers [79]. Each codebook has eight multidimensional complex codewords that correspond to eight points of constellation. The length of each codeword is four, which is the same as the spreading length. Upon transmission, the codeword of each layer is selected based on the input bit sequence. The codewords from different layers are overlaid with each other.

Figure 5.27 summarizes the main features of SCMA, and the explanations are listed as follows:

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

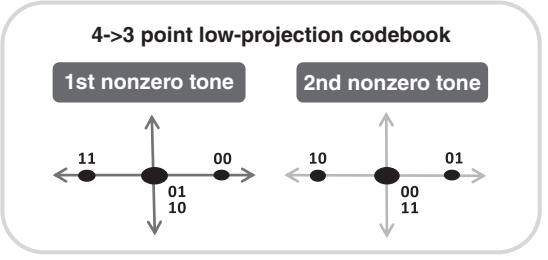


Figure 5.28 Example of low-projection SCMA codebook.

- **Code domain signal superposition**: SCMA allows superposition of multiple symbols from different users on each resource element (RE). For example, on subcarrier 1, symbols from UE1, 3, and 5 are overlapped with each other. The superposition pattern on each RE can be different and is defined in the SCMA codebook.
- **Sparse spreading**: To reduce interlayer interference so that more symbol collisions can be tolerated with low receiver complexity. This allows overloading, meaning that more data layers than the spreading length can be accommodated. For example, in Fig. 5.27, six data layers can be supporting by spreading length 4, resulting in an overloading factor of 150.
- **Multidimensional constellation**: For better SE and low receiver complexity.
- **SCMA codebook design**: Codebook design is the key feature that distinguishes SCMA from other NoMA schemes. The design of the SCMA codebook can be considered as the joint optimization of the sparse spreading pattern design and the multidimensional constellation design. In general, the aim of the codebook design is to provide good distance properties (Euclidean and/or product) among the points in the overall multidimensional constellation to maximize the coding/shaping gain. Another feature of SCMA codebooks is the possibility of having a lower number of projection points over each resource element. This is due to the multi-dimensional nature of the codebooks, which allows two constellation points to collide over some of the nonzero components, as they can still be separated over the other nonzero components. An example is shown in Fig. 5.28, in which the constellation points corresponding to 01 and 10 collide over the first tone, but are separated over the second tone, making three projection points instead of four [79]. This feature can be considered in the design of SCMA codebooks with the goal of reducing the receiver complexity.
- **Receiver**: For a non-orthogonal system like SCMA, there are more than one OFDM symbols overlaid on each RE, so joint multiuser detection algorithms are needed. In general, maximum a posteriori probability (MAP) detection is optimal but with very large complexity. Furthermore, due to the sparsity of the SCMA codeword structure, a message-passing algorithm (MPA) on a factory graph with much lower complexity can be adopted to achieve a suboptimal performance. Although a MPA has significantly lower complexity over MAP detection, it is still

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

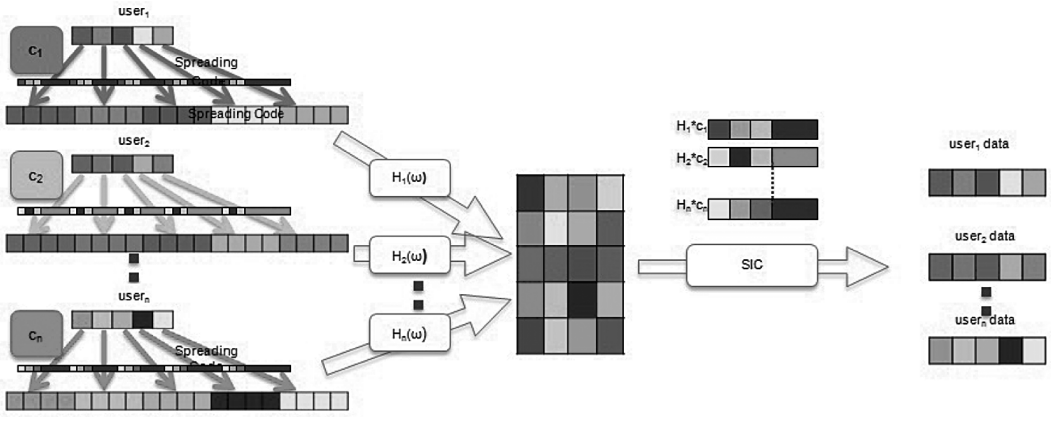


Figure 5.29 Concept of multi-user shared access (MUSA).

suffering from implementation complexity and a wide dynamic range of exponential operations when computing the likelihood function. Then, inspired by the idea of max-log-MAP decoding algorithm for turbo codes, a further simplification can be achieved by implementing MPA in a log domain. This subsection is about the implementation of MPA in a log domain, which is named as log-MPA. By log-domain transform, exponential operations are omitted, multiplication operations are replaced by addition operations, and addition operations are replaced by maximum operations.

MUSA

Multiuser shared access (MUSA) is a NoMA scheme operating in the code domain. Conceptually, each user's modulated data symbols are spread firstly by a specially designed sequence that facilitates robust successive interference cancellation (SIC) implementation compared to the sequences employed by traditional DS-CDMA (direct-sequence CDMA). Then, each user's spread symbols are transmitted concurrently on the same radio resource by means of shared access, which is essentially a superposition process. Finally, decoding each user's data from a superimposed signal can be performed at the BS side using SIC technology. The major processing blocks of the MUSA transmitter and receiver are illustrated in Fig. 5.29 [79].

The spread sequence design is a key component of MUSA, and it has a direct impact on the system performance and computation complexity of the corresponding SIC implementation. Long pseudo-random spread sequences used by traditional DS-CDMA, such as in IS-95 standard, may exhibit relatively low cross-correlation even if the number of sequences is larger than the length. Thus, those sequences can provide a soft capacity limit on the system rather than a hard capacity limit. This soft capacity limit concept can also be understood as the overloading ability of a system. Long spread sequences may be attractive in terms of soft capacity limit, however, the SIC receiver tends to be less efficient when very large spreading factors are used and the system

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

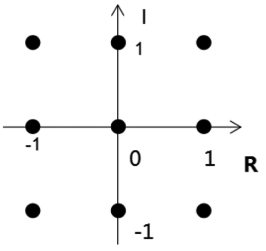


Figure 5.30 Constellation of elements of MUSA sequences.

needs to be operated in extremely overloaded situations to achieve a reasonably good capacity. MUSA relies on a special family of complex spread sequences that can enjoy relatively low cross-correlation even when they are very short, e.g., eight, or even four. In one example of MUSA spread sequence, the real and imaginary parts of the complex spread sequence are from an M-ary real value set. By this method, even the short spread sequences with real and imaginary part selected from a simple 3-value set, $-1,0,1$, can deliver quite impressive performance in terms of overloading. The corresponding trilevel constellation is depicted in Fig. 5.30 [79].

It should be pointed out that the spread sequences used in MUSA are different from the spreading codes in the sense that MUSA spreading does not have the low density property. While the low density codes are more friendly to advanced symbol-level detectors such as using a MPA, the codeword-level SIC can downplay the importance or necessity of using the advanced detectors. Equipped with the well-optimized spreading sequence and state-of-the-art SIC technology, MUSA is capable of decoupling the multiuser mingled data, even if those users are contending to access the system. Potentially a large number of devices are allowed to transmit data at their will (by randomly picking spread sequences) spread the data, and send them. In another words, MUSA is suitable for the scenario where the UL transmissions are not tightly scheduled, and the grants for transmission are not signaled on a per-user basis, and with a high overloading. The relaxed UL synchronization requirement for MUSA allows simple derivation of UL time from a DL synchronization process, which can greatly cut down on battery consumption. Lastly, the code domain superposition nature of MUSA can turn the near–far problem into a near–far advantage. The disparity in the received signal-to-noise ratio (SNR) across simultaneously transmitting users can be exploited in MUSA to facilitate SIC. Tight transmit power control is no longer needed, which can further lower the device cost and its power consumption.

PDMA

PDMA (pattern division multiple access) is a kind of NoMA technology based on the principle of the introduced reasonable diversity between multiuser to promote the capacity, which can obtain higher multiuser multiplexing and diversity gain by designing a multiuser diversity PDMA pattern matrix to implement non-orthogonal signals transmission in such domains as time, frequency, power, and space. PDMA can design patterns for specific users in time, frequency, and space resources. Figure 5.31 [79]

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

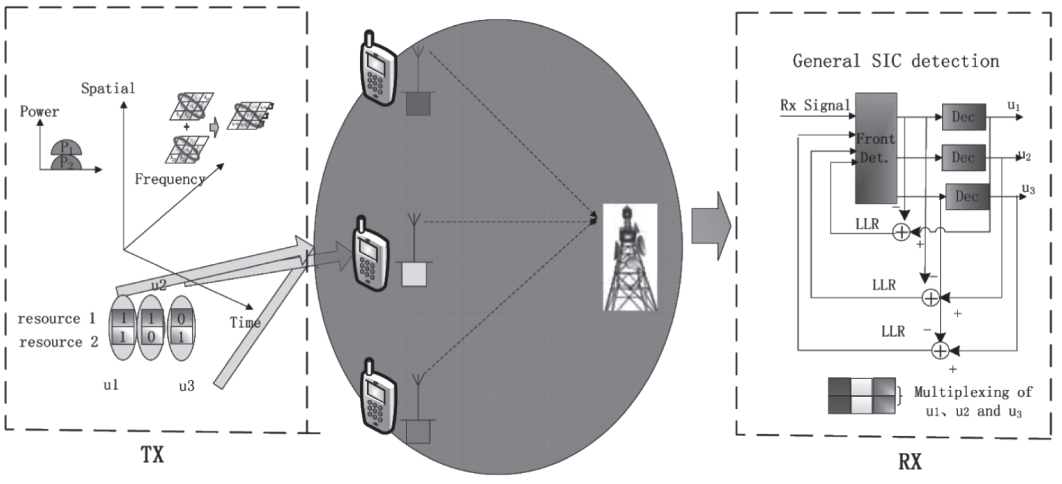


Figure 5.31 The technical framework of the PDMA UL application.

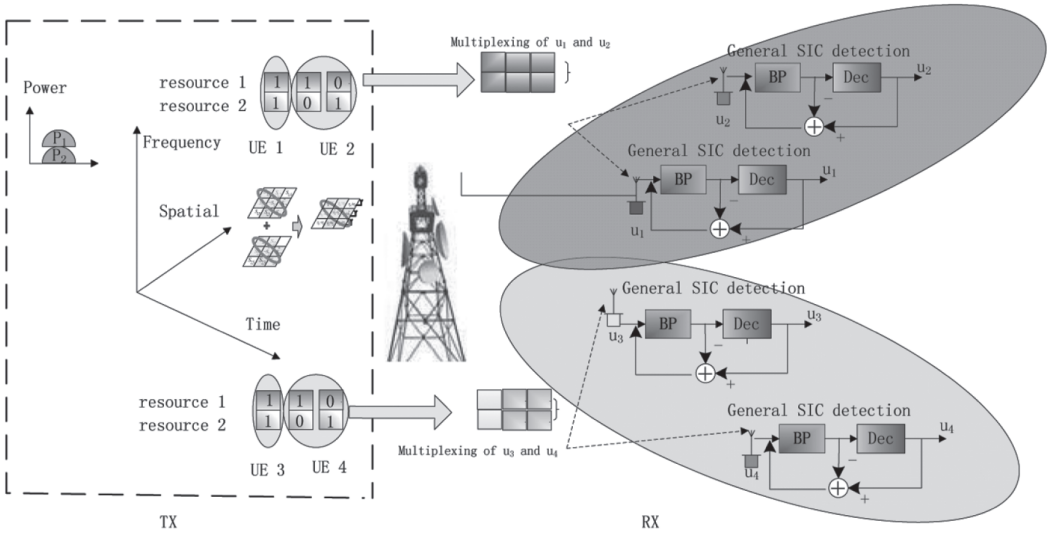


Figure 5.32 The technical framework of the PDMA DL application.

shows the technical framework of the PDMA UL application, Fig. 5.32 [79] shows that of the PDMA DL application.

As shown in Figs. 5.31 and 5.32, the PDMA technical framework includes two parts, the transmitter and the receiver, which reflects that the PDMA technology consider the joint design of the transmitter and the receiver, based on the optimization point of view for a multiuser communication system. On the transmitter side, we distinguish users by using the non-orthogonal characteristic pattern based on the multiple signals domain (including time, frequency and the space domain, etc.). On the receiver side, we can realize suboptimal multiuser detection by general SIC, based on the features of the user pattern.

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

5.6.2 Theoretical Analysis of a NoMA System

Constellation-Constrained (CC) Capacity

In this section, to provide insight into the achievable sum rate for NoMA in the UL, we analyze the constellation-constrained (CC) capacity [80] of NoMA schemes in the multiple access channel (MAC). The CC capacity is measured by the mutual information between the input and the output in a Rayleigh fading channel, where modulated symbols of each user are constrained to a finite set of constellation points with a uniform distribution. For a K-user MAC channel, the received symbol vector $\mathbf{y}$ at the base station is

$$\mathbf{y} = \mathbf{H} \odot \mathbf{S}\mathbf{x} + \mathbf{n} = \mathbf{H}_{\mathit{eff}}\mathbf{x} + \mathbf{n} \tag{5.21}$$

where $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_K]$ of size $N \times K$ denotes the channel matrix for all K users, where the (i, j)-th entry $h_{i,j}, \forall i \in \{0, 1, \cdots, N-1\}$, is assumed to be an independent and identically distributed (i.i.d) complex Gaussian random variable with zero mean and unit variance, $\mathbf{H}_{\mathit{eff}} = \mathbf{H} \odot \mathbf{s}$ denotes the effective channel matrix for all k users, $\mathbf{x} = [x_1, x_2, \cdots, x_K]^T$ refers to the transmitted symbol vector of all k users with normalized power $\mathrm{E}\left[|x_i|^2\right] = 1$, $\odot$ denotes the element-wise Hadamard product of two matrices, and finally $\mathbf{n} \sim \left(0, \sigma^2\mathbf{I}\right)$ is the noise vector.

For illustrative purposes, we utilize sparse code-based NoMA to exemplify the NoMA scheme. We write the received signal vector y for the sparse code-based NoMA scheme with sparse pattern matrix $\mathbf{P}_{2\times3} = \begin{bmatrix} 1 & \sqrt{2} & 0 \\ 1 & 0 & \sqrt{2} \end{bmatrix}$ as follows:

$$\mathbf{y} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \end{bmatrix} \odot \begin{bmatrix} 1 & \sqrt{2} & 0 \\ 1 & 0 & \sqrt{2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \mathbf{n} = \mathbf{H}_{\mathit{eff}}\mathbf{x} + \mathbf{n}. \tag{5.22}$$

Using the chain rule from the information theory, we can express the sum of the CC capacity as follows:

$$\mathrm{I}(\mathbf{x}, \mathbf{y}) = \mathrm{I}(x_1; \mathbf{y}) + \mathrm{I}(x_2; \mathbf{y}\,|x_1) + \mathrm{I}(x_3; \mathbf{y}\,|x_1, x_2)\,. \tag{5.23}$$

The term $\mathrm{I}(\mathbf{a}; \mathbf{b})$ denotes the mutual information between the variables $\mathbf{a}$ and $\mathbf{b}$, whereas the term $\mathrm{I}(\mathbf{a}; \mathbf{b}\,|\mathbf{c})$ denotes the mutual information between the variables $\mathbf{a}$ and $\mathbf{b}$, conditioned on the knowledge of the variable $\mathbf{c}$. Where $\mathrm{I}(x_1; \mathbf{y}) = \mathrm{H}(\mathbf{y}) - \mathrm{H}(\mathbf{y}\,|x_1)$, $\mathrm{H}(\mathbf{y}) = -\int p(\mathbf{y})\log_2(p(\mathbf{y}))d\mathbf{y}$, $p(\mathbf{y}) = \frac{1}{\Pi_{i=1}^{3}|\chi_i|}\sum_{\chi} p(\mathbf{y}\,|\mathbf{x})$, and χ_i denotes the size of modulation order of the k-th user, which is assumed to be 4 (i.e., QPSK constellation is considered) for all users in this paper without loss of generality. To realize 150% overloading, without loss of generality, we can also utilize the following sparse pattern matrix for an example and others are not precluded.

$$\mathbf{P}_{4\times6} = \begin{bmatrix} \sqrt{2} & \sqrt{2} & \sqrt{2} & 0 & 0 & 0 \\ \sqrt{2} & 0 & 0 & \sqrt{2} & 0 & \sqrt{2} \\ 0 & \sqrt{2} & 0 & \sqrt{2} & \sqrt{2} & 0 \\ 0 & 0 & \sqrt{2} & 0 & \sqrt{2} & \sqrt{2} \end{bmatrix} \tag{5.24}$$

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

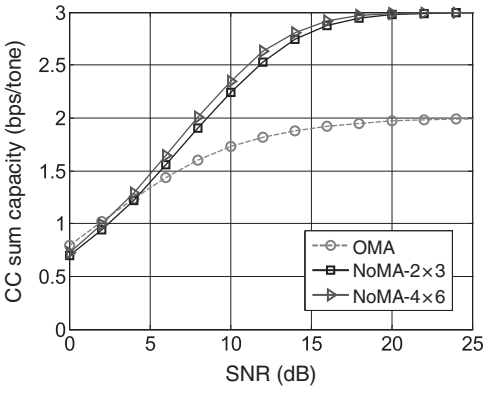


Figure 5.33 Numerical results of constellation constraint capacity for UL sparse code-based NoMA overloading 150%.

Then, we can compute the CC sum capacities of these different NoMA schemes with different sparse pattern matrices in Fig. 5.33, where the CC sum capacity of OMA is also shown for comparison.

Here we want to emphasize that the goal of our comparison for the UL MA study is the UL access user number for a given system target spectrum efficiency (the range of operation point under the given modulation order), not the single-user throughput. So the same modulation order (QPSK) is assumed for each user both in the OMA case and NoMA case in the above numerical simulation.

5.6.3 A Unified Framework of Multiple Access Schemes

Advanced MA technology has been envisioned as one of key enablers of 5G communications. The signals from different users will be superposed into the same time and frequency resource and demodulated by an advanced receiver algorithm to provide higher spectrum efficiency and system capability. Grant-free transmission will be allowed to significantly reduce signaling overhead, shorten access latency, and decrease terminal power consumption. The MA techniques as introduced in the literature are summarized in Table 5.6.

The just-discussed advanced MA schemes, as well as the traditional OMA scheme, e.g., OFDMA, are both identified as potential candidates for 5G. Based on the diverse deployment scenarios and traffic requirements of 5G, flexible MA can be utilized to meet the verified demands. For example, in the case of massive connections, the question of how to accommodate more users with limited resources has become a critical problem for next-generation access networks. With NoMA schemes, e.g., SCMA, MUSA, PDMA, or RSMA [81], the same resources are shared and reused by multiple users, thus the number of connections increases. To support the traffic with low latency requirement, NoMA schemes help to realize grant-free MA, with which the latency is much lower, and the power consumption of the devices can be reduced. In other scenarios, such as DL machine-type traffic, the simple OMA schemes are better, due

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

Table 5.6 Summary of multiple access techniques.

	BDM	MUSA	SPC-NOMA	PDMA	RSMA	SCMA
Scenario	DL eMBB	UL MMC, DL eMBB	eMBB, MMC, URC	eMBB, MMC, URC	UL MMC/ UL URC	eMBB, MMC, URC
Multiplexing domain	Code/ Power	Code/ Power	Power	Code/Power /Spatial	Code/ Power	Code/ Power
Transmitter Overloading	High	High	Medium	High	High	High
Transmitter Spreading	No	Yes	No	Yes	Yes	Yes
Transmitter multidimensional constellation	No	No	No	No	No	Yes
Receiver	MMSE/SIC	SIC	SIC	SIC/MPA	SIC	MPA/SIC
Receiver Complexity	Low (SSD), Medium (MSD)	Medium	Medium	Medium	Medium	Medium

to device cost and implementation complexity. A study of scenarios and requirements for next-generation access technologies has been made. The requirement of support for wide a range of services to be deployable on a single continuous block of spectrum in an efficient manner is proposed in the document. To support this operational requirement, we propose a compatible MA structure as depicted in Fig. 5.34 [82]. By the unified structure, we can flexibly configure different MA schemes according to various 5G scenarios.

Without loss of generality, here we take the DL transmission as an example, the unified MA framework can also be used in the UL transmission. As depicted in Fig. 5.34, the differences among various MA schemes lie in the different realization of bit-level and symbol-level's operations, e.g., cell-specific/user-specific interleaver design in the

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

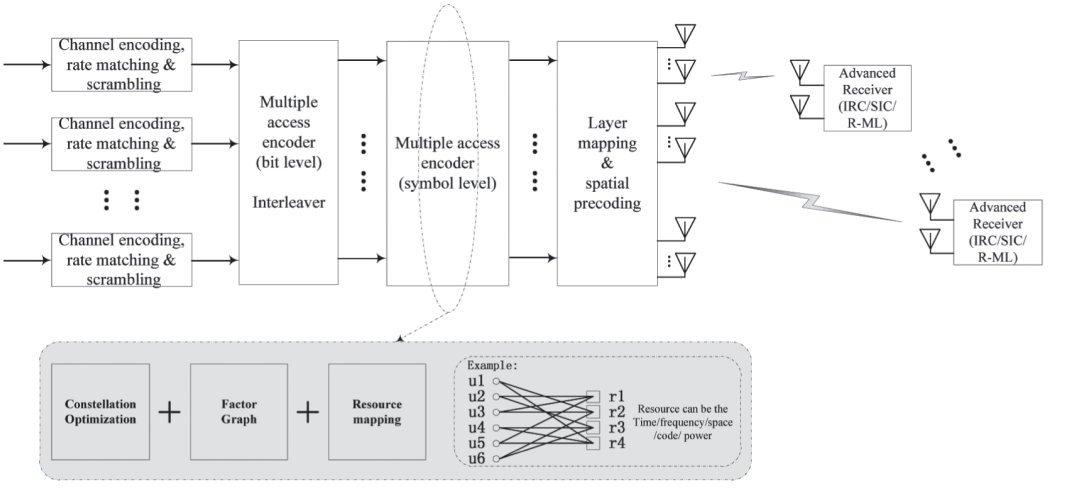


Figure 5.34 Unified framework of MA schemes.

bit-level processing, constellation optimization, factor graph, and multiplexing domain in the symbol-level operations. The detailed explanations are listed in Table 5.7.

Note 1: The identity/permutation matrix means independent/dependent mapping method respectively. For example, like the MUST Cat-1 interleaver, the coded bits of each user is mapped to symbols without considering the bit/symbol of the other co-scheduled user, such mapping is independent mapping. Otherwise it is dependent mapping (permutation matrix in interleaver).

Note 2: The constraint permutation matrix means that it is not a totally dependent mapping method. The bits or symbols that are mapped should satisfy certain conditions, e.g., bits from the same user should be adjacent.

Note 3: The element of the sparse matrix maybe different among various schemes, e.g., the element of the sparse matrix can only be "0" or "1" in the SCMA and PDMA , and it can also be "−1" in the MUSA.

5.6.4 Summary

NoMA is an attractive solution to boost system capacity by accommodating more users at the same time/frequency resource, and reducing system latency caused by scheduling and queuing, as well as relaxing the dependency on precise channel state information and feedback quality. In particular, for UL, NoMA-enabled grant-free is a competitive solution for small packet transmission in many scenarios, including mMTC, URLLC, and eMBB.

In the coming study of 3GPP NoMA SI, more works will be dedicated to the comprehensive evaluations of the various candidate schemes based on the unified framework to better understand the commonality and differentiation of different schemes, to find the recommended configurations for different target scenarios. Moreover, as other technologies are evolving in parallel in 3GPP, the study of how these radio technologies can be

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

Table 5.7 Configuration method of different MA schemes based on the unified framework.

		Interleaver	Constellation mapping	Factor graph	Resource mapping (multiplexing domain)
OMA		Identity matrix[1]	Gray-mapped legacy constellation	Identity matrix	time/frequency/code/space
NOMA	MUST Cat 1	Identity matrix	non-Gray-mapped superposed constellation	Identity matrix	power
	MUST Cat 2	Constraint permutation matrix[2]	Gray-mapped superposed constellation	Identity matrix	power/bit
	MUST Cat 3	Permutation matrix	Gray-mapped legacy constellation	Identity matrix	bit
	SCMA	Identity matrix	joint optimization (multidimensional modulation + Sparse matrix[3])		code/power
	PDMA	Identity matrix	legacy modulation	Sparse matrix	code/power/space
	… …	… …	… …	… …	… …

integrated with NoMA shall be carried out. As one example, the integration of NoMA with (massive) MIMO has been raised in the literature. In addition, the efficient and flexible MA adaptation schemes, the MA codebook and RS design, and low-complexity receiver design are also key to the commercialization of NoMA in the future. The EE performance at the UE side is an important KPI. Therefore, the EE–SE codesign of various MA schemes needs to be investigated in-depth [1].

5.7 Full Duplex

Out of the potential technologies for 5G and beyond, full duplex has drawn much attention because it may bring a revolution of the duplex mode in future wireless

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

communications. The current TDD mode is half duplex, since its DL and UL are on the same frequency but not simultaneous. The FDD mode is a full duplex mode in time domain by transmitting and receiving simultaneously, but operates on different frequencies in the DL and UL. Transmitting and receiving essentially at the same time and frequency, full duplex may maximally achieve doubled SE compared with either TDD or FDD. Much progress has been made so far in the research and development of full duplex, with multiple successful demos of the feasibility of short-range wireless connection in either relay or single access point scenarios [83–90].

One fundamental challenge to full duplex is self-interference cancellation, the self-interference with the following techniques prescribed to solve this challenge: antenna cancellation, analog cancellation, and digital cancellation. As the antenna number increases at the access point, the self-interference cancellation will become more complicated, since an analog cancellation circuit is generally needed for each Tx and Rx antenna pair, to cancel possible multipath self-interference signals. Currently the analog cancellation circuit is large, e.g., the prototype board measures 10×10cm for a single Tx and single Rx full duplex transceiver [90]. More efforts are still required in the design of efficient and space-compact interference cancellation circuits.

Some solutions have been proposed to realize full duplex with multiple antennas. One solution is to utilize antenna cancellation via symmetric placement of multiple Tx and Rx antennas where the performance was better than conventional MIMO, but required N radios and 2N antennas [87]. Digital signal processing techniques can further mitigate self-interference in MIMO full duplex systems, e.g., the time-domain transmit beamforming method [88]. When the antenna array becomes a massive array with over 64 elements, the excessive degrees of freedom in massive MIMO full duplex systems can be leveraged [89]. When the antenna size in the MIMO full duplex system is a problem, reducing the Tx and Rx antenna number can significantly reduce the array size. A full duplex implementation using a single antenna was presented in [90], where novel analog and digital cancellation techniques were utilized to cancel the self-interference to the receiver noise floor. In addition to the self-interference cancellation, MAC mechanisms were also investigated. For example, the full duplex physical layer was designed in [91] with a MAC protocol backward compatible with current IEEE 802.11 systems.

In the case of a multicell full duplex network, interference management becomes even more complex [92]. Recently, full duplex networking issues were investigated when full duplex is considered for a wireless network. An interference management strategy was proposed in [93] to handle the intercell interference to achieve gains in data rates over half duplex systems, with the assumption that all BSs in the network have instantaneous access to the global CSI. Extensive system-level simulations were carried out in [94] to evaluate the throughput of full duplex cellular systems, where a suboptimal resource allocation scheme was considered. The results showed that full duplex could significantly increase the aggregate throughput of current cellular systems in both DL and UL. However, these analyses and simulations may be overly optimistic, since instantaneous availability of the CSI and perfect interference mitigation are assumed.

The EE performance of a full duplex system has also been investigated. In [95], the authors analyzed SE–EE trade-off with a full-duplex BS and half-duplex UE under

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

two different kinds of residual self-interference. Corresponding optimization algorithms were proposed to achieve EE maximization. [96] studied four types of power control schemes in full-duplex networks, which shows remarkable gains on EE compared to half-duplex networks. In [97], joint beamforming and time allocation algorithms were proposed to maximize the sum rate and EE, taking account of energy harvest-enabled UE. EE maximization in full-duplex networks with MIMO techniques were studied in [98, 99]. In [98], precoding in the context of full-duplex MU-MIMO was studied, and in [99] power allocation algorithm was studied in the context of full-duplex same cell with massive MIMO. In [100], the authors investigated critical EE challenges in implementing full-duplex relaying in mm-Wave systems and outlined a number of promising EE-oriented solutions for designing full-duplex relaying-enabled systems.

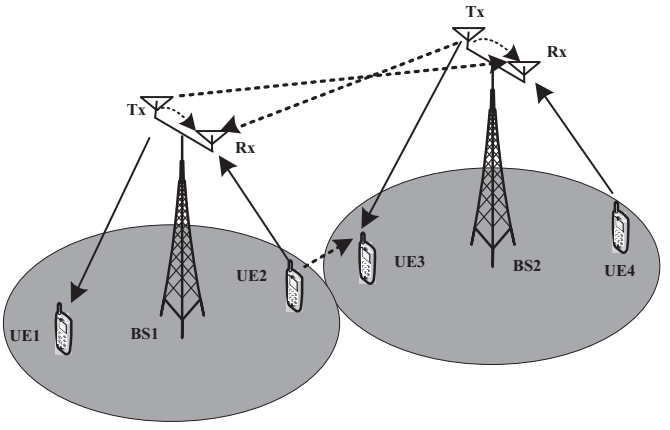
In a practical system deployment, severe intercell and intracell interferences due to simultaneous transmission and reception in each cell make the deployment of full duplex networks very difficult. For example, DL/UL channel measurement and estimation may not be easily achieved due to mutual interference from both the same cell and adjacent cells. In addition, proper scheduling of DL and UL users requires inter-user channel information, which also may incur heavy signaling overhead. Though tremendous progress has been made in the study and implementation of full duplex technologies, there exist many open issues for successful deployment of full duplex network in 5G and beyond, including, for example: a desirable full duplex frame structure and the required modification of the current standards, DL and UL reference signals, efficient intracell/inter-cell interference mitigation, transceiver structures, the extension of TDD and FDD to full duplex, and implementation of MIMO full duplex.

This section aims to shed some insights on how a full duplex cellular network should be designed, from the perspectives of interference mitigation techniques, frame structure design, and TDD/FDD extension to full duplex. So far, 5G NR has started to standardize the flexible frame structure with configurable DL and UL transmissions, and the mechanism of cross-link interference mitigation. There are still many open issues pending in the research and standardization. The work in this section is expected to provide some reference designs.

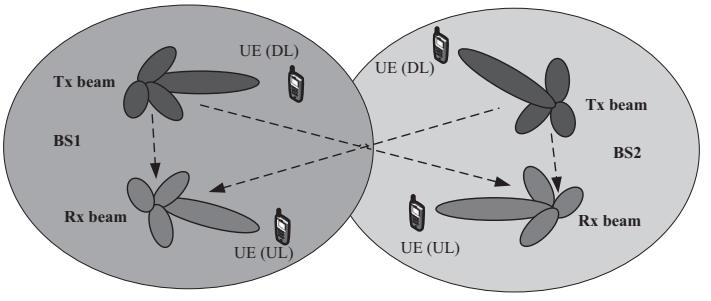
5.7.1 Interference Mitigation in Full Duplex Networks

In the current TDD or FDD system, the DL-to-DL interference received at the UE and the UL-to-UL interference received at the BS have been extensively studied in literature and standardization bodies. For example, the CoMP technologies were standardized in 4G LTE-A and IEEE 802.16m to counteract these co-channel interferences. In a multicell full duplex network, the interference condition is more severe. A two-cell full duplex network is shown in Fig. 5.35a, where BS1 and BS2 are transmitting to UE1 and UE3 in the DL, respectively, while UE2 and UE4 are transmitting to BS1 and BS2 in the UL, respectively. In addition to the self-interference from the Tx to Rx at each BS, there are intracell UL-to-DL interferences from UE2 to UE1 and from UE4 to UE3, intercell UL-to-DL interferences from UE2 to UE3, and DL-to-UL interferences from BS1 Tx

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press



(a) Interferences in full duplex network



(b) Interference mitigation via beamforming

Figure 5.35 Interferences mitigation in a full duplex system.

to BS2 Rx and from BS2 Tx to BS1 Rx. These interferences have significant impact on whether a full duplex system works, and must be mitigated properly, in addition to the already existing DL-to-DL and UL-to-UL interferences.

In the multicell scenario, due to the interferences explained above, the DL and UL in each cell should be jointly scheduled for an optimized performance (e.g., maximum sum rate of both DL and UL). Self-interference mitigation, which is the predominant issue in the point-to-point application of full duplex, should be jointly considered with many other issues, like DL and UL user pairing, DL and UL power control, DL-to-UL and UL-to-DL (either intracell or intercell) interference mitigation, and the DL and UL QoS.

UL-to-DL Interference Channel Measurement and Feedback

To maximize the sum rate of both DL and UL in a full duplex system, minimizing UL-to-DL interference is very important. This requires that the BS schedules proper DL and UL users with negligible inter-user interference. In the following, an intracell

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

UL-to-DL interference channel measurement and feedback scheme is presented, which can be extended to the intercell case easily. The key features of this scheme are listed below:

- Simultaneous transmission of orthogonal DL and UL RS: The DL and UL RSs should be orthogonal (e.g., in code or in frequency domain) and transmitted simultaneously, such that the DL UEs can measure the DL and UL RSs simultaneously and calculate the SINR accurately.
- Orthogonal UL RSs: The BS allocates orthogonal RSs resource to each prescheduled UL UE, then each UL UE will transmit its own RSs with a given transmit power, e.g., the max power. The number (N) of the UL UEs, and the UL RSs are broadcasted.
- Feedback mechanism: The DL UEs feedback the lowest M ($M \leq N$, determined by the scheduler) interference power and the corresponding UL RS indexes. If some UL RSs are not received due to large propagation attenuation between the DL and UL UEs, the associated interference power is assumed to be 0. The indexes of these un-received UL RSs are also known to the DL UEs since N and all the UL RSs are broadcasted.

It should be pointed out that large-scale, fading-based approaches can be utilized alternatively to reduce the possibly increased overhead of the above scheme, e.g., the DL and UL UEs with a sufficiently large gap between their large-scale fading to the BS can be scheduled.

Joint Interference Mitigation

In addition to the UL-to-DL interference information, the BSs also need to know all the channel information: the DL and UL channels of each UE with the adjacent BSs, the self-interference channels at BSs, and the DL-to-UL interference channels. Note that the self-interference channels and the DL-to-UL interference channels are semi-static, and hence allow much lower overhead in channel estimation.

As shown in Fig. 5.35b, one feasible approach to maximizing the sum rate is via joint Tx and Rx beamforming [92, 101] when multiple antennas are available at the BSs. The DL data of each BS is precoded such that DL data transmission is improved while the self-interference, the DL-to-UL interferences, and the DL-to-DL interferences are mitigated, e.g., nulls are formed toward the corresponding Rx antennas. To further improve performance, UL beamforming can be adopted similarly.

The above joint interference mitigation approach requires instantaneous CSI at the BSs and a central controller responsible for all the corresponding signal processing. This can be significantly facilitated via the C-RAN [102]. With well-designed DL and UL RSs for various measurement purposes, full or partial CSI of all users are possibly available at the C-RAN baseband pool. System level optimization can be made possible in a real sense.

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

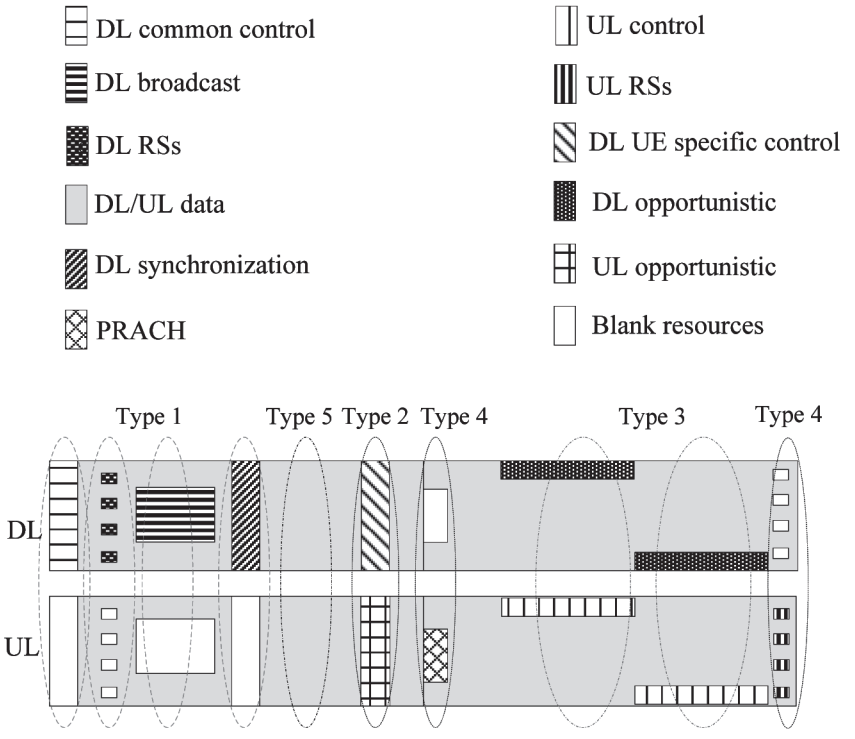


Figure 5.36 Frame structure for full duplex.

5.7.2 Full Duplex Frame Structure Design

Obviously, the TDD frame structure is not applicable directly to full duplex, since it does not support simultaneous DL and UL transmissions. Neither is the FDD frame structure, since applying the DL and UL frame structures of the FDD system (e.g., FDD-LTE) on the same frequency may bring severe interference between DL and UL signals, e.g., between the UL PRACH and the PDCCH (which is a kind of DL common control channel). Therefore, the full duplex frame structure is required to both support simultaneous DL and UL transmissions and to have the capability of DL and UL interference mitigation. This is a dramatic departure from conventional TDD and FDD systems and brings unique technical challenges.

As a virtual bridge between the cellular network and the mobile users, OTA signaling and control information are essential to the stability, reliability, and operation efficiency of the system. Naturally, the basic principle of full duplex frame structure design is that the DL (UL) signaling and control information have a higher priority than data and hence should not be interfered by UL (DL) data, signaling, and control. As one potential example, high-level DL and UL frame structures are illustrated in Fig. 5.36,

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

with the following five types of regions to cover almost all the important DL and UL transmissions of signaling, control, and data.

Type 1: DL Signaling and Control with UL Being Muted

In this region, the essential DL signaling and control information are transmitted, the transmission in the corresponding UL resources need to be muted to ensure that no DL UE may fail to detect these DL signals due to interferences from UL UEs. This information includes the following:

- Primary synchronization sequence (PSS) and secondary synchronization sequence (SSS), which are DL synchronization signals to enable UEs to identify the presence of a network, and acquire the time and frequency of the network.
- The main information block (MIB) in physical broadcast channel (PBCH), which contains the most essential physical-layer information like system bandwidth, frame number, etc., and must be detected successfully before further access to more information in the system information block (SIB).
- Cell common control signals, which deliver control messages to support radio resource management and data transmissions. A UE needs to decode the following three DL control channels before it receives and decodes the data on the PDSCH allocated to it.
- The PDCCH, which provides physical-layer signaling to support MAC-layer operations. The common PDCCH, which carries information such as paging, PRACH response, and system information, belongs to region type 1. On the contrary, the user-specific PDCCH does not belong to this region.
- The physical hybrid-ARQ indicator channel (PHICH), which includes the information of HARQ ACK/NACK feedback for UL data transmissions on the PUSCH.
- The physical control format indicator channel (PCFICH), which indicates the number of OFDM symbols designated for PDCCHs in the current subframe.
- Cell common DL RSs, which are designed for CSI measurement or estimation.

Type 2: DL Signaling and Control with Opportunistic UL

In this region, the UE-specific control information, or the non-delay-sensitive radio resource management signaling is transmitted on the PDSCH. The UL transmission is possible only if its interference to the DL UEs is tolerable. The UL is therefore opportunistic.

Type 3: UL Signaling and Control with Opportunistic DL

In this region, the UL SRS, physical UL control channel (PUCCH) and DMRS are transmitted. The transmission of the DL data or UE specific control information is possible if the UL-to-DL interference can be managed by scheduling proper DL and UL users. The DL is therefore opportunistic.

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

Type 4: UL Signaling and Control with DL Being Muted

- When the UE transmits the random access signal on the PRACH for network entry, the DL transmission is difficult to be scheduled since the UL UEs transmitting PRACH are not known to the BS yet. Therefore, the DL is suggested to be muted.
- When some special UL control signals are transmitted, the DL transmission needs to be muted. For example, when the UL channels of one UE to adjacent BSs need to be measured accurately for intercell interference mitigation, DL transmissions at the adjacent BSs should be muted to avoid interference to the UL reception at each BS (e.g., the Tx of BS2 is muted when UE2 transmits RSs for channel measurement to BS1 and BS2, as shown in Figure 5.35a.

Type 5: Opportunistic DL/UL Data

For data transmission, the BS can freely schedule DL and UL users.

In addition to the interference-aware considerations above, another issue that needs to be addressed is the time gap between DL and UL in a traditional TDD system. This time gap was designed to be sufficiently large to provide enough time for the DL-to-UL switch of the BS circuitry, and to mitigate intracell and intercell DL and UL interferences. In a full duplex system, this gap can be removed because the BS is not expected to quickly switch from DL to UL, since Tx and Rx have separate radios. In addition, fast switching in the UE from DL to UL is not necessary, because the UE's DL and UL can be scheduled with a time gap.

With the above design principles, the full duplex frame structure can be devised, e.g., based on either a TDD-LTE or FDD-LTE frame structure. More efforts are needed to elaborately design each type of region.

5.7.3 Extension of FDD and TDD to Full Duplex

Potential standardization of full duplex key technologies in LTE-A or IEEE 802.11ax may start in the near future. The focus may lie in leveraging full duplex capabilities at infrastructure nodes to support half duplex UEs, since full duplex UEs still seem impractical due to, e.g., complexity and the large size of the current analog cancellation circuit.

Extending the traditional FDD system (e.g., operating on the carrier frequency f1 in the UL and f2 in the DL) to a full duplex system generally requires doubled transceiver number for both f1 and f2. As shown in Fig. 5.37a, the full duplex BS (with f1 and f2 in both DL and UL) transmits to FDD UE1 on f2, and receives from UE1 on f1 simultaneously. This is exactly the FDD mode to UE1. Meanwhile, the BS transmits to FDD UE2 on f1 and receives from UE2 on f2. This is also the FDD mode to UE2. However, with the full duplex frame structure in Fig. 5.36 on both f1 and f2, the current FDD UEs cannot be supported directly, thus mandating corresponding changes to the UE design. Moreover, the design of UE1 and UE2 should be different, since UE1 is operating on f1 in the UL and f2 in the DL, while UE2 is operating on f1 in the DL and f2 in the UL.

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

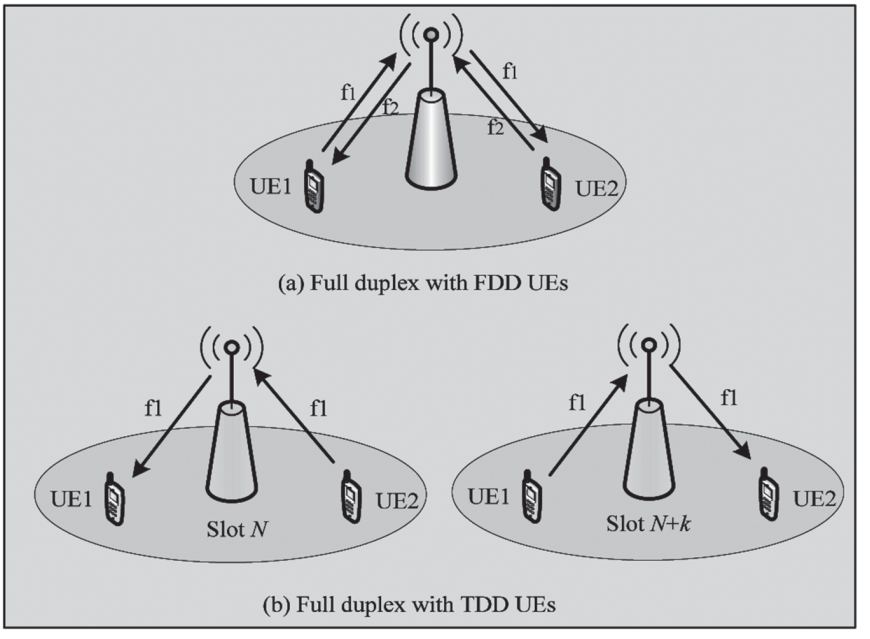


Figure 5.37 TDD and FDD extension to full duplex.

The extension of the traditional TDD system (e.g., operating on the carrier frequency f1 in both DL and UL) to full duplex is shown in Fig. 5.37b. The BS schedules TDD UE1 in the DL and TDD UE2 in the UL in time slot N, and schedules UE1 in the UL and UE2 in the DL in time slot $N + k$ (k is an integer determined by the BS scheduler). The current TDD UEs still operate in TDD mode, while the whole cell operates in the full duplex mode. Different from the FDD case, the current TDD UEs can be supported directly by the full duplex BS with frame structure in Fig. 5.36.

5.7.4 Summary

In this section, several key design issues for full duplex networks were investigated, with potential solutions proposed. Firstly, the interference situation was analyzed and potential interference mitigation techniques were discussed. Design principles of the full duplex frame structure were then presented, aiming to provide efficient means to mitigate the severe interferences in full duplex networks. The extension of traditional TDD and FDD systems to full duplex was further addressed.

Full duplex technologies are expected to reduce both the control and data-plane latency and to double the link and system capacity maximally via simultaneous DL and UL operation. The application of full duplex in future wireless communication systems like 5G is able to remove the clear distinction between TDD and FDD and to better utilize the unpaired spectrum. With growing interest and efforts on the research and development of full duplex technologies, especially on fundamental issues like

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

efficient intracell/intercell interference mitigation, full duplex frame structure, DL and UL reference signals, implementation of full duplex with multiple antennas, etc., full duplex network can be feasibly deployed in the future.

Recall that three types of frame structures are discussed in Subsection 5.3.1 for the DL, UL, and bidirectional transmission, respectively. They are fundamentally different from the full duplex frame structure presented in this section. The bidirectional transmission within one subframe of the self-contained frame structure is time duplexed, not full duplex in essence. The framework of SDAI allows flexible configuration of frame structures to enable real full duplex. However, more work is to be done in the future standardization of full duplex, based on the 5G NR specification on frame structure.

5.8 Flexible Signaling, Control, and Protocol

5.8.1 Introduction

Considering the intensified diversity of user traffic, conventional network signaling, control and protocol (SCP) are facing tremendous challenges, especially in scenarios with critical performance targets in 5G networks.

The conventional "one-size-fits-all" mechanism offers undifferentiated network signaling, control, and protocol toward all kinds of mobile traffic. Take LTE for example: there are two RRC states for the user, i.e., RRC_IDLE and RRC_CONNECTED. A UE in the RRC_IDLE state maintains no connection with the network. It is free from frequent interactions with the network, thus is energy efficient. On the other hand, a UE in the RRC_CONNECTED state is connected to the network and is required to maintain the connection. When a RRC_IDLE UE wants to transmit data, it is required to first establish an RRC connection by executing the RRC connection setup procedure. Then, after expiry of a data-inactive timer, the RRC connection between the UE and the network is released for terminal power savings. As demonstrated in table 5.2.1–2 in [103], each RRC connection setup/release procedure requires about 18 signaling interactions with a 265-bytes payload. Such RRC procedure incurs very heavy signaling overhead in small packet transmission [104], and therefore is not energy efficient. Meanwhile, by the LTE procedure, the control-plane latency, which refers to the time to move from a battery efficient state (e.g., IDLE) to start of continuous data transfer (e.g., ACTIVE), can hardly be satisfied, considering its 10ms target in 5G. As a result, to optimize network performance and efficiency for each individual 5G scenario, scenario- and service-aware flexible SCP design is motivated.

With the flexible SCP framework, different SCP function components (e.g., UE states and scheduling mechanisms) can be orchestrated by the RRC and radio resource management (RRM) entity for different traffic scenarios. For example, for eMBB, in the control plane, novel RRC state and RRC procedures could be designed to enable fast connection establishment; in the user plane, more dedicated treatment could be applied to differentiate user traffic and enhance the user experience. For mMTC of massive small data connectivity, slim signaling can be introduced in both the control plane and

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

user plane to reduce signaling consumption. For URLLC, radio resources could be pre-allocated to reduce data latency incurred by resource grant.

The flexible SCP is an element of the SDAI slice. In the later part of this chapter, new SCP function components are introduced: firstly, a new RRC state with lean signaling design is introduced for mMTC and eMBB for control-plane latency reduction and signaling overhead reduction; Secondly, grant-free MAC scheduling is introduced for mMTC signaling reduction and URLLC low-latency data transmission; Thirdly, the concept of smart RAN is introduced to enhance users' experience with optimized cross-layer processing for different services.

5.8.2 New SCP Function Components

Service-Oriented New UE State: RRC_KEEP_ALIVE / RRC_INACTIVE

In [104] it is revealed that small data bursts result in orders of magnitude higher OTA signaling overhead than the streaming services, by a metric termed as data-signaling ratio (DSR). As a solution, a new UE state, RRC_KEEP_ALIVE is proposed to support both low signaling overhead and small packet transmission. UE behaviors of RRC_KEEP_ALIVE are characterized as follows:

1. No RRC connection maintenance;
2. No handover;
3. Context reservation and slim signaling before data transmission;
4. Small data transmission;

Since the RRC_KEEP_ALIVE state requires no RRC connection maintenance, the UE in RRC_KEEP_ALIVE behaves like an RRC_IDLE UE and will consume much reduced power. Meanwhile, since only slim or little signaling is needed before data transmission, state transition signaling can be saved and fast data transmission is enabled.

It is evaluated that, for a UE with periodical keep-alive data transmission, application of RRC_KEEP_ALIVE achieves a 6-fold DSR gain compared with the conventional RRC_IDLE/CONNECTED approach. This is because the RRC_KEEP_ALIVE state saves RRC maintenance signaling compared to the RRC_CONNECTED state, and it requires less signaling for activation of data transmission compared to the RRC_IDLE state.

In 3GPP Release 15, a similar concept, RRC_INACTIVE, is introduced. This standardized new UE state is mainly motivated by the requirements for fast RRC state transition and reduced control plane latency. It is clearly characterized in [105] by the following features:

1. Broadcast of system information;
2. Cell reselection mobility;
3. 5G core and NG-RAN connection (both control and user planes) is established for UE;
4. The UE access stratum context is stored in at least one gNB and the UE;

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

5. Paging is initiated by NG-RAN;
6. Discontinuous reception (DRX) for NG-RAN paging configured by NG-RAN;
7. RAN-based notification area (RNA) is managed by NG-RAN;

Obviously, though the two states are somehow different in motivations, they do share most behaviors and merits. For both of the states: UE context is reserved and slim signaling can be enabled; UE behaves similar as in energy saving RRC_IDLE state. RRC_INACTIVE complements the RRC_KEEP_ALIVE state with the feature of accurate paging support in RAN. And RRC_KEEP_ALIVE is keener on optimization of small data users, which supports small data transmission without RRC state transition.

Service- and Load-Aware MAC Scheduling: From Grant-Based to Grant-Free

To further resolve the issues of low signaling efficiency (e.g., low DSR) and long data transmission latency in small data transmission, grant-free scheduling is investigated in 5G NR. By pre-allocating radio resources to grant-free users via RRC signaling, grant-free scheduling enables users to transmit UL data without further resource request and grant. In this way, grant signaling in the MAC layer can be saved. Data transmission latency could possibly be reduced, since the user data could be transmitted before a dedicated grant is possibly received by the UE.

The performance difference between the grant-free and the grant-based scheduling is shown in Fig. 5.38, where the system bandwidth is 1.08MHz, and the basic scheduling resource unit is 180kHz in frequency and 0.5ms in time. Suppose in every 10 TTIs, there is a dedicated TTI configured for the small data. Then there are N=1200 basic scheduling units. Assume the system serves 300,000 static mMTC users, each with a traffic arrival rate of 1 packet per 5 minutes. In this simulation, the grant-free transmission is free of any signaling interaction, while the grant-based transmission involves signaling of scheduling request and scheduling grant. For the grant-based approach, the total bandwidth is divided into two parts: one for scheduling request, and the other for data transmission and accompanying grant signaling. The resource of scheduling request is configured similar to PRACH in LTE, i.e., 12 basic resource units for 64 preambles and each preamble conveys a scheduling request. For each grant, overhead of 3/7 scheduling resource unit is consumed. It is assumed that the SE is 1.8 and 1 bps/Hz for the grant-based and grant-free UEs, respectively.

A metric termed “effective spectral efficiency” is introduced for evaluation. It takes into account the signaling impacts on data transmission efficiency and is calculated as the total transmitted data payload bits divided by total radio resources consumed by both data payload and supporting signaling per Hz per second.

As can be seen in Fig. 5.38, efficiency of the grant-based scheduling is quite resistant to load increase, and it improves with increased packet sizes. This explains why grant-based scheduling is widely adopted in existing cellular systems. However, it is also observed that efficiency of grant-based scheduling is low in the case of small data packets, while the grant-free solution is the other way around. Obviously, grant-free is especially efficient to small data transmission and may serve as an ideal solution for the

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

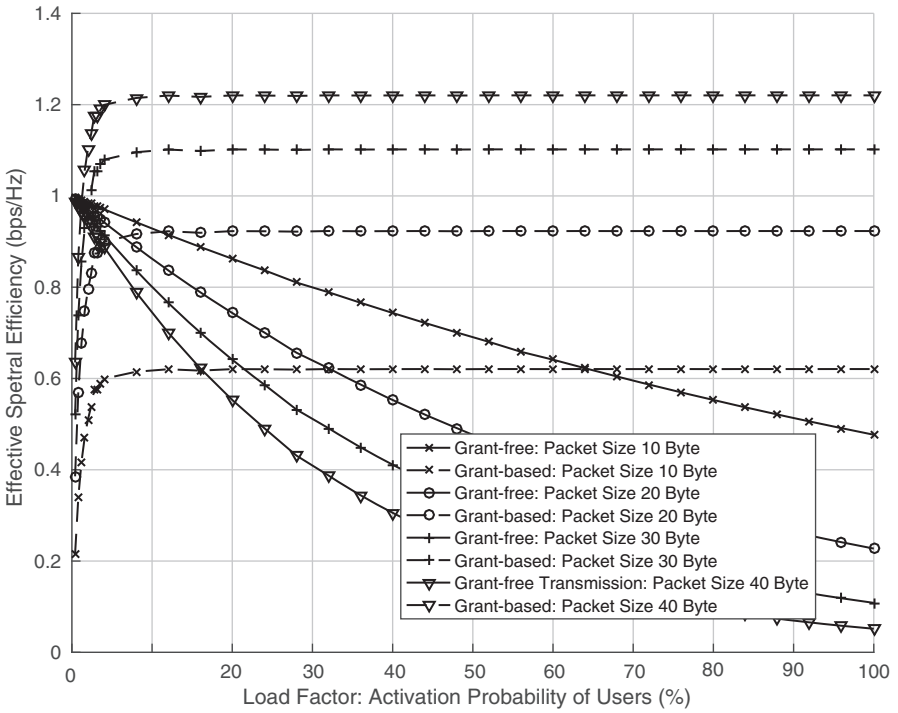


Figure 5.38 Comparison of grant-free and grant-based MAC scheduling.

5G mMTC scenario for signaling and energy consumption reduction. Besides, it can be also applied to the latency stringent URLLC services. Therefore, grant-free scheduling can be taken as a new and important component of the 5G MAC scheduler.

In ongoing 3GPP standardization, the 3GPP Release 15 is devoted to study grant-free solutions by RRC preconfiguration of grant-free resources. The pre-allocated grant-free resources may be shared by one or a group of UEs.

Context-Aware Service Delivery: Smart RAN

Given the ever-increasing mobile traffic load and limited radio resources, it would be very challenging to meet the requirements of mobile applications in 5G. To make it worse, the limited resources in even the current mobile networks are far from being fully utilized. One major cause lies in the isolated design of the mobile network and mobile applications. The mobile network radio resource allocation is performed based on rapidly varying radio conditions, while rate adjustment of the application is based on relatively long-term statistical E2E throughput observations. Therefore, there is a clear mismatch between the air interface data rate and the application data rate, which may lead to inefficient resource usage and degraded user experience. Operators would be highly motivated to remove the separation wall between the mobile network and mobile application.

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

Smart RAN is a novel concept proposed by CMCC. It is dedicated to enable more real-time coordination between RAN and the applications. On the one hand, smart RAN is capable of mobile application awareness, and is able to configure RAN strategy accordingly. Its awareness of applications can be assisted by the application layer. For example, considering the DL video, the video application server can mark the video packets with different labels, e.g., primary video segment packets or supplementary packets. Upon detection of different labels, RAN configures differentiated radio protocol stack for different packets. For example, a higher MAC scheduling priority may be applied to packets with a "primary" label. Meanwhile, the video client in UE can report to RAN its application buffer status, based on which RAN allocates preemptive radio resource to the buffer-hungry UEs. On the other hand, smart RAN is capable of exposing the RAN condition (e.g., the available radio bandwidth) as required by the application server via e.g., application programmable interfaces (APIs) to the video server. Then, the video server can adjust the video bit rate according to the radio bandwidth. By alignment of the network and the application, the network resources can be more fully utilized and the UE experience can be improved. Furthermore, to enable more timely coordination between RAN and the applications, edge deployment of the application server can be considered in smart RAN.

Both the industry and academia have been devoting efforts to smart RAN case studies. In a Release 14 study item led by CMCC, "Context-aware service delivery," application use cases of smart RAN are discussed, including data caching, TCP data transmission, and video transmission [106]. The simulation results in its annex show that "video playout buffer-aware scheduling" achieves a 25% capacity gain and significantly reduces the stalling probability. In [107], a cross-layer moving mean algorithm (CMMA) is proposed; it is demonstrated that by radio condition-aware application rate adjustments, the CMMA can boost throughput of the typical video protocol "dynamic adaptive streaming" over HTTP (DASH) by up to 30%. The reason is that the video client in DASH conventionally chooses an optimum video code rate by estimation of available bandwidth based on throughput observations on the client side. Nevertheless, since time granularity of the client observations is much larger than the millisecond-level radio variations, video transmission can hardly fully utilize the available radio bandwidth. With assistance of the radio throughput information provided by RAN, the DASH server is capable of estimating the throughput condition more accurately.

5.8.3 Summary

In this section, some examples of new SCP function components were introduced to meet the 5G requirements. First, new RRC_KEEP_ALIVE/RRC_INACTIVE states in the RRC layer were introduced, which can be easily merged for the satisfaction of both signaling reduction for mMTC and control-plane latency reduction for eMBB. Then, grant-free scheduling in the MAC layer was proposed to avoid dedicated resource grant signaling for massive small packets, and to reduce URLLC data latency. Furthermore, smart RAN was proposed to overcome the gap between the network and applications, and to achieve more efficient radio resource usage. It is possible to further distinguish

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

traffic types within both a session and a bearer, and to apply finer optimization accordingly. Flexible SCP, as a soft and green solution, is expected to be implemented in 5G mobile networks and beyond.

References

[1] Q. Sun, S. Han, C.-L. I, and Z. Pan, "Software defined air interface: A framework of 5G air interface," *IEEE WCNC*, 2015.

[2] C.-L. I, S. Han, Z. Xu et al. "New paradigm of 5G wireless internet," *IEEE J. on Sel. Areas in Commun.*, vol. 34, no. 3, pp. 472–482, Mar. 2016.

[3] S. Zhang, G. Wang, and C.-L. I, "Is mmWave ready for cellular deployment," *IEEE Access*, 2017, DOI: 10.1109/ACCESS.2017.2711491.

[4] R. Crane, *Propagation Handbook for Wireless Communication System Design*, CRC Press, 2003.

[5] A. F. Molisch, *Wireless Communications*, Second Edition, John Wiley & Sons, 2012.

[6] T S. Rappaport, *Wireless Communications: Principles and Practice*, Second Edition, Prentice Hall, 2002.

[7] J. Ramakrishna, *Radiowave Propagation and Smart Antennas for Wireless Communications*, Springer, 2001.

[8] T. L. Marzetta, *Fundamentals of Massive MIMO*, Cambridge University Press, 2016.

[9] B. M. Hochwald, T. L. Marzetta, and V. Tarokh, "Multiple-antenna channel hardening and its implications for rate feedback and scheduling," *IEEE Trans. on Inf. Theory*, vol. 50, no. 9, pp. 1893–1909, Sept. 2004.

[10] H. Q. Ngo and E. G. Larsson, "No downlink pilots are needed in TDD massive MIMO," *IEEE Trans. on Wireless Commun.*, vol. 16, no. 5, pp. 2921–2935, May 2017.

[11] K. Zheng, S. Ou, and X. Yin, "Massive MIMO channel models: A survey," *International Journal of Antennas and Propagation*, vol. 2014.

[12] L. Liu et al., "The COST 2100 MIMO channel model," *IEEE Commun. Mag.*, vol. 19, no. 6, pp. 92–99, Dec. 2012.

[13] 3GPP, TR 36.873, "Study on 3D channel model for LTE (Release 12)."

[14] J. Meinilä, P. Kyösti, L. Hentilä et al., "Deliverable 5.3: WINNER+ final channel models," WINNER+/Celtic project CP5-026, Jun. 2010.

[15] S. Salous, "COST IC1004 white paper on channel measurements and modeling for 5G networks in the frequency bands above 6 GHz," 2016. www.ic1004.org/.

[16] ETSI, "New ETSI group on millimetre wave transmission starts work," 2015. www.etsi.org/news-events/news/866-2015-01-press-new-etsigroup-on-millimetre-wave-transmission-starts-work.

[17] 3GPP, TR 38.900, "Study on channel model for frequency spectrum above 6 GHz (Release 14)," 2016.

[18] Various Contributors, "White Paper on 5G channel model for bands up to 100 GHz," white paper, 2015. www.5gworkshops.com/5G_Channel_Model_for_bands_up_to100_GHz(2015-12-6).pdf.

[19] S. Jaeckel, L. Raschkowski, K. Borner, and L. Thiele, "QuaDRiGa: A 3-D multi-cell channel model with time evolution for enabling virtual field trials," *IEEE Trans. Antenna Propag.*, vol. 62, no. 6, pp. 2921–2935, Jun. 2014.

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

[20] IEEE Wireless LAN medium access control (MAC) and physical layer (PHY) specifications amendment 3: Enhancements for very high throughput in the 60 GHz band," 2012. https://ieeexplore.ieee.org/document/6392842.
[21] MiWEBA, "Channel modeling and characterization," June 2014. www.miweba.eu/wp-content/uploads/2014/07/MiWEBA_D5.1_v1.011.pdf.
[22] METIS, "METIS Channel Models," 2015.
[23] mmMAGIC, "Measurement campaigns and initial channel models for preferred suitable frequency ranges" 2016. https://5g-mmmagic.eu/results/deliverables.
[24] M. K. Samimi and T. S. Rappaport, "3-D millimeter-wave statistical channel model for 5G wireless system design," *IEEE Trans. on Microwave Theory and Techniques*, vol. 64, no. 7, pp. 2207–2225, Jul. 2016.
[25] X. Lin, J Andrews et.al, "An overview of 3GPP device-to-device proximity services," *IEEE Commun. Mag.*, vol. 52, no. 4, pp. 40–48, Apr. 2014.
[26] X. Cheng, Y. Li, B. Ai, et al., "Device-to-device channel measurements and models: a survey," *IET Communications*, 2014.
[27] R. He, et al., "High-Speed Railway Communications: From GSM-R to LTE-R," *IEEE Veh. Technol. Mag.*, vol. 11, no. 3, pp. 49–58, Sept. 2016.
[28] B. Ai, et.al., "Challenges toward wireless communications for high-speed railway," *IEEE Trans. on Intell. Transp. Syst.*, vol. 15, no. 5, pp. 2143–2158, Oct. 2014.
[29] R.. He et al., "A measurement based stochastic model for high speed railway channels," *IEEE Trans. on Intell. Transp. Syst.,* vol. 16, no. 3, pp. 851–854, Jun. 2015.
[30] G. Li, B. Ai, K. Guan et al., "Path loss modeling and fading analysis for channels with various antenna setups in tunnels at 30 GHz band," *IEEE EuCAP*, 2016.
[31] T. Zhou, C. Tao, S. Salous, and L. Liu, "Measurements and analysis of angular characteristics and spatial correlation for high-speed railway channels," *IEEE Trans. on Intell. Transp. Syst.*, vol. 19, no. 2, pp. 357–367, Feb. 2018.
[32] 3GPP RAN1 R1-162156, "Scenario & design criteria on flexible numerologies," Huawei, HiSilicon.
[33] Q. Li, G. Li, W. Lee et al., "MIMO techniques in WiMAX and LTE: A feature overview," *IEEE Commun. Mag.*, vol. 48, no. 5, pp. 86–92, 2010.
[34] 3GPP, TR 38.913, "Study on scenarios and requirements for next generation access technologies."
[35] F. Rusek, D. Persson, B. K. Lau et al., "Scaling up Mimo: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan 2013.
[36] S. Han, C.-L. I, Z. Xu, and C. Rowell, "Large-scale antenna systems with hybrid analog and digital beamforming for millimeter wave 5G," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 186–194, Jan. 2015.
[37] Z. Xu, S. Han, Z. Pan, and C.-L. I., "Alternating beamforming methods for hybrid analog and digital MIMO transmission," *2015 IEEE Int. Conf. on Commun. (ICC)*, pp. 1595–1600.
[38] W. Roh, J. Seol et al., "Millimeter-wave beamforming as an enabling technology for 5G cellular communications: Theoretical feasibility and prototype results," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 106–113, Feb. 2014.
[39] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 101–107, Jun. 2011.
[40] F. Khan, Z. Pi, and S. Rajagopal, "Millimeter-wave mobile broadband with large scale spatial processing for 5G mobile communication," *50th Annual Allerton Conf. on Commun., Control, and Comput. (Allerton)*, 2012, pp. 1517–1523.

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

[41] O. E. Ayach, R. W. Heath, S. Rajagopal, and Z. Pi, "Multimode precoding in millimeter wave MIMO transmitters with multiple antenna sub-arrays," *2013 IEEE Global Commun. Conf. (GLOBECOM)*, pp. 3476–3480.

[42] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath Jr.,"Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

[43] S. Kutty and D. Sen, "Beamforming for millimeter wave communications: An inclusive survey," *IEEE Commun. Surveys & Tutorials*, vol. 18, no. 2, pp. 949–973, 2016.

[44] X. Gao, L. Dai, S. Han, C.-L. I, and R. W. Heath, "Energy-efficient hybrid analog and digital precoding for mmWave MIMO systems with large antenna arrays," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 998–1009, Apr. 2016.

[45] A. Mezghani, F. Antreich, and J. A. Nossek, "Multiple parameter estimation with quantized channel output," *Smart Antennas (WSA)*, pp. 143–150.

[46] L. Jacques, J. N. Laska, P. T. Boufounos, and R. G. Baraniuk, "Robust 1-Bit compressive sensing via binary stable embeddings of sparse vectors," *IEEE Trans. on Inf. Theory*, vol. 59, no. 4, pp. 2082–2102, Apr. 2013.

[47] J. Mo and R. W. Heath, "Capacity analysis of one-bit quantized MIMO systems with transmitter channel state information," *IEEE Trans. Signal Process.*, vol. 63, no. 20, pp. 5498–5512, Oct. 2015.

[48] A. Liu and V. Lau, "Sum capacity of massive MIMO systems with quantized hybrid beamforming," in *2016 IEEE Int. Symposium on Inf. Theory (ISIT)*, pp. 320–324, 2016.

[49] T. Shin, G. Kim, H. Park, and H. M. Kwon, "Quantization error reduction scheme for hybrid beamforming," *18th Asia-Pacific Conf. on Commun. (APCC)*, pp. 243–247, 2012.

[50] T. Demir and T. E. Tuncer, "Hybrid beamforming with two bit RF phase shifters in single group multicasting," *2016 IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, pp. 3271–3275.

[51] H. Q. Ngo, E. G. Larsson, and T. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. on Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.

[52] Z. Xu, Z. Pan, and C.-L. I, "Fundamental properties of the EE-SE relationship," *IEEE Wireless Commun. and Netw. Conf. (WCNC)*, 2014.

[53] G. Y. Li, Z. Xu, C. Xiong et al., "Energy-efficient wireless communications: Tutorial, survey, and open issues," *IEEE Wireless Commun.*, vol. 18, no. 6, pp. 28–35, Dec. 2011.

[54] 3GPP, "Final report of 3GPP TSG RAN WG1 #85," 2016.

[55] 3GPP, "Final report of 3GPP TSG RAN WG1 #86," 2016.

[56] 3GPP, "Final report of 3GPP TSG RAN WG1 #86b," 2016.

[57] Qualcomm, "OFDM based waveform single user evaluation," R1-164685, Nanjing, China, May 2016.

[58] J. Abdoli, M. Jia, and J. Ma, "Filtered OFDM: A new waveform for future wireless systems," *IEEE 16th Int. Workshop on Signal Process. Advances in Wireless Commun. (SPAWC)*, 2015.

[59] Huawei, "f-OFDM scheme and filter design," R1-164033, Nanjing, China, May 2016.

[60] X. Wang, T. Wild, F. Schaich, A. Santos, and Alcatel-Lucent, "Universal filtered multi-carrier with leakage-based filter optimization," *European Wireless*, 2014.

[61] B. Farhang-Boroujeny, "OFDM versus filter bank multicarrier: development of broadband communication systems," *IEEE Signal Proc. Magazine*, pp. 92–112, May 2011.

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

[62] M. Bellanger, "FS-FBMC: an alternative scheme for filter," *ISCCSP*, Rome, Italy, May 2012.
[63] N. Michailow, M. Matth, I. S. Gaspar et al., "Generalized frequency division multiplexing for 5th generation cellular networks," *IEEE Trans. on Commun.*, vol. 62, no. 9, Sept. 2014.
[64] A. Farhang, N. Marchetti, L. E. Doyle, "Low Complexity transceiver design for GFDM," 2015. arXiv:1501.02940v1.
[65] Cohere Technologies, "OTFS modulation waveform and reference signals for new RAT," R1-162930, Busan, South Korea, Apr. 2016.
[66] A. M. Sayeed and B. Aazhang, "Joint multipath-doppler diversity in mobile wireless communications," *IEEE Trans. on Commun.*, vol. 47, no. 1, pp. 123–132, Jan. 1999.
[67] F. Hasegawa, S. Shinjo, A. Okazaki et al., "Static sequence assisted out-of-band power suppression for DFT-s-OFDM," *PIMRC 2015*, Hong Kong, pp. 61–65, Sept. 2015.
[68] G. Berardinelli, F. M. L. Tavares, T. B. Sorensen et al., "Zero-tail DFT-spread-OFDM signals," *2013 IEEE Global Commun. Conf. (GLOBECOM)*, Sept. 2013.
[69] D. A. Guimaraes, "Contributions to the understanding of the MSK Modulation," *REVISTA Telecommunications*, vol. 11, no. 1, Dec. 2008.
[70] K. Murota and K. Hirade, "GMSK modulation for digital mobile radio telephony," *IEEE Trans. on Commun.*. vol. 29, no. 7, pp. 1044–1050, Jul. 1981.
[71] 3GPP RANI #86 R1-167963, "Way forward on waveform," Huawei, Aug. 2016.
[72] 3GPP #84-BIS R1-163222, "Waveform for NR," Ericsson, Apr. 2016.
[73] 3GPP #37 R1-040642, "Comparison of PAR and cubic metric for power de-rating," Motorola, May 2004.
[74] 3GPP RANI #86-BIS R1-1609929, "Discussion and evaluation of UL waveforms," CMCC, Oct. 2016.
[75] 3GPP RANI #85 R1-166004, "Response LS on realistic power amplifier model for NR waveform evaluation," RAN4, Nokia.
[76] 3GPP TS 36.101, "User Equipment (UE) radio transmission and reception (Release 12)," 2016.
[77] 3GPP TR 36.873, "Study on 3D channel model for LTE," 2014.
[78] 3GPP TR 38.913, "Study on scenarios and requirements for next generation access technologies."
[79] "White paper, v2.0D-5G enabler: Alternative multiple access v1," Nov. 2015. http://www.future-forum.org/2009cn/member.asp.
[80] R1-164889, 3GPP WG1, "Analytical evaluation of multiple access and preliminary LLS results," CMCC. http://www.3gpp.org.
[81] R1-163510, 3GPP WG1, "Candidate NR multiple access schemes," Qualcomm. http://www.3gpp.org.
[82] R1-162870, 3GPP WG1, "On unified framework for multiple access schemes," CMCC. http://www.3gpp.org.
[83] S. Hong et al, "Applications of self-interference cancellation in 5G and beyond," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 114–121, Feb. 2014.
[84] DUPLO Deliverable D2.1, "Design and measurement report for RF and antenna solutions for self-interference cancellation," Apr. 2014.
[85] M. Duarte and A. Sabharwal, "Full duplex wireless communications using off-the-shelf radios: Feasibility and first results," *Proc. IEEE Asilomar Conf. on Signals, Systems and Computers*, pp. 1558–1562 Nov. 2010.

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

[86] M. Jain, J. I. Choi, D. Bharadia et al., "Practical, real-time, full duplex wireless," *Proc. ACM 17th Annual Int. Conf. on Mobile Comput. and Networking (MobiCom)*, pp. 301–312, Aug. 2011.

[87] E. Aryafar, M. A. Khojastepour, K. Sundaresan, S. Rangarajan, and M. Chiang, "Midu: enabling MIMO full duplex," in *Proc. ACM 17th Annual Int. Conf. on Mobile Comput. and Networking (MobiCom)*, pp. 257–268, Aug. 2012.

[88] Y. Hua, P. Liang, Y. Ma, A. C. Cirik, and Q. Gao, "A method for broadband full duplex MIMO radio," *IEEE Signal Process. Lett.*, vol. 19, no. 12, pp. 793–796, Dec. 2012.

[89] B. Yin, M. Wu, C. Studer, J. R. Cavallaro, and J. Lilleberg,, "Full duplex in large-scale wireless systems," in *Proc. IEEE Asilomer Conf. on Signals, Systems, and Computers*, Nov. 2013.

[90] D. Bharadia, E. McMilin, and S. Katti, "Full duplex radios," in *Proc. Sigcomm*, Aug. 2013.

[91] B. P. Day, A. R. Margetts, D. W. Bliss, and P. Schniter, "Full duplex bidirectional MIMO: Achievable rates under limited dynamic range," *IEEE Trans. Signal Process.*, vol. 60, no. 7, pp. 3702–3713, Jul. 2012.

[92] DUPLO Deliverable D4.1.1, "Performance of Full duplex systems," Jan. 31, 2014.

[93] A. Sahai, S. Diggavi, and A. Sabharwal, "On uplink/downlink full duplex networks," *Proc. IEEE Asilomer Conf. on Signals, Systems, and Computers*, Nov. 2013.

[94] S. Goyal, P. Liu, S. Hua, and S. Panwar, "Analyzing a full duplex cellular system," *Proc. 47th IEEE Annual Conf. on Inf. Sciences and Systems (CISS)*, pp. 1–6, Mar. 2013.

[95] D. Wen et al., "Results on energy- and spectral-efficiency tradeoff in cellular networks with full-duplex enabled base stations," *IEEE Trans. on Wireless Commun.*, vol. 16, no. 3, pp. 1494–1507, Mar 2017.

[96] C. Feng et al., "Power control in full duplex networks: Area spectrum efficiency and energy efficiency," *IEEE Int. Conf. Computers (ICC)*, 2017.

[97] V. Nguyen et al., "Spectral and energy efficiencies in full-duplex wireless information and power transfer," *IEEE Trans. on Commun.*, vol. 65, no. 6, pp. 2220–2223, May 2017.

[98] D. Nguyen et al., "Precoding for full duplex multiuser MIMO systems: Spectral and energy efficiency maximization," *IEEE Trans. on Signal Process.*, vol. 61, no. 16, pp. 4038–4051, Aug. 2013.

[99] Y. Li et al., "On the spectral and energy efficiency of full-duplex small-cell wireless systems with massive MIMO," *IEEE Trans. on Veh. Technol.*, vol. 66, no. 3, pp. 2339–2353, Mar. 2017.

[100] Z. Wei et al., "Energy-efficiency of millimeter-wave full-duplex relaying systems: Challenges and solutions," *IEEE Access*, vol. 4, pp. 4848–4860, 2016.

[101] D. Nguyen, L.-N. Tran, P. Pirinen, and M. Latva-aho, "On the spectral efficiency of full duplex small cell wireless systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 9, pp. 4896–4910, Sept. 2011.

[102] C.L. I, J. Huang et al, "Recent progress on C-RAN centralization and cloudification," IEEE Access, vol. 2, pp. 1030–1039, Aug. 2014.

[103] 3GPP TR 36.822, "LTE radio access network (RAN) enhancements for diverse data applications (Release 11)."

[104] Y. Chen, G. Li, Z. Pan, and C.-L. I, "Small data optimized radio access network signaling/control design," *IEEE Int. Conf. on Commun.*, pp. 49–54, 2014.

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press

[105] 3GPP TS 38.300 "NR; Overall description; Stage 2 (Release 15)."
[106] 3GPP TR 36.933 "Study on context aware service delivery in RAN for LTE (Release 14)."
[107] H. Liu, Y. Liu, Z. Chen, L. Sang, and D. Yang, "A cross-layer bandwidth estimation algorithm for DASH services," *Proc. IEEE Int. Carnahan Conf. Security Technol. (ICCST)*, 2015.

https://doi.org/10.1017/9781108277389.006 Published online by Cambridge University Press