1 Overview of New Technologies for 5G Systems

Vincent W. S. Wong, Robert Schober, Derrick Wing Kwan Ng, and Li-Chun Wang

1.1 Introduction

In recent years, wireless service providers in different countries have deployed both the Third Generation Partnership Project (3GPP) Long Term Evolution (LTE) and LTE-Advanced systems. Despite the unprecedented data rates and quality of service (QoS) provided by these new networks, user demand is beginning to exceed their capabilities. For example, the proliferation of smartphones and tablets has caused a significant and sustained increase in mobile data traffic. In fact, in 2015 alone, global mobile data traffic grew by 74% from 2.1 to 3.7 exabytes [1]. Furthermore, the existing networks are not well suited for the exceedingly large number of devices and appliances that are expected to be connected wirelessly to the Internet in future Internet of Things (IoT) applications and machine-to-machine (M2M) communications. Moreover, to make the growth of networks and the number of connected devices economically and ecologically sustainable, energy efficiency has to be substantially improved. Also, emerging new applications such as remote surgery in healthcare, autonomous driving, and wireless control of industrial robots require ultra-low latencies in the sub-millisecond range and ultra-high reliability, giving rise to the notion of the Tactile Internet. In order to support the exponential growth of existing mobile traffic and the emergence of new wireless applications and services, researchers and standardization bodies worldwide have set out to develop a fifth generation (5G) of wireless networks [2–6]. Some of the stringent requirements for this next generation of wireless networks are listed in Table 1.1 [7].

To meet these challenging requirements, a mere evolution of the current networks is not sufficient. Instead, a true revolution of technologies in both the radio access network and the mobile core network is needed (Figure 1.1). In the radio access network, fundamentally new physical layer technologies such as massive multiple-input multiple-output (MIMO), non-orthogonal multiple access (NOMA), full-duplex (FD) communication, millimeter wave (mmWave) communication, device-to-device (D2D) communication, and visible light communication (VLC) will be deployed. Furthermore, leveraging cloud computing, the cloud radio access network (C-RAN) has emerged as a promising and cost-efficient mobile network architecture to enhance the spectrum and

Table 1.1 Requirements for 5G wireless communication systems [7].

Figure of merit	5G requirement	Comparison with 4G
Peak data rate	10 Gb/s	100 times higher
Guaranteed data rate	50 Mb/s	–
Mobile data volume	10 Tb/s/km^2	1000 times higher
End-to-end latency	Less than 1 ms	25 times lower
Number of devices	1 M/km^2	1000 times higher
Total number of human-oriented terminals	$\geq$ 20 billion	–
Total number of IoT terminals	$\geq$ 1 trillion	–
Reliability	99.999%	99.99%
Energy consumption	–	90% less
Peak mobility support	$\geq$ 500 km/h	–
Outdoor terminal location accuracy	$\leq$ 1 m	–

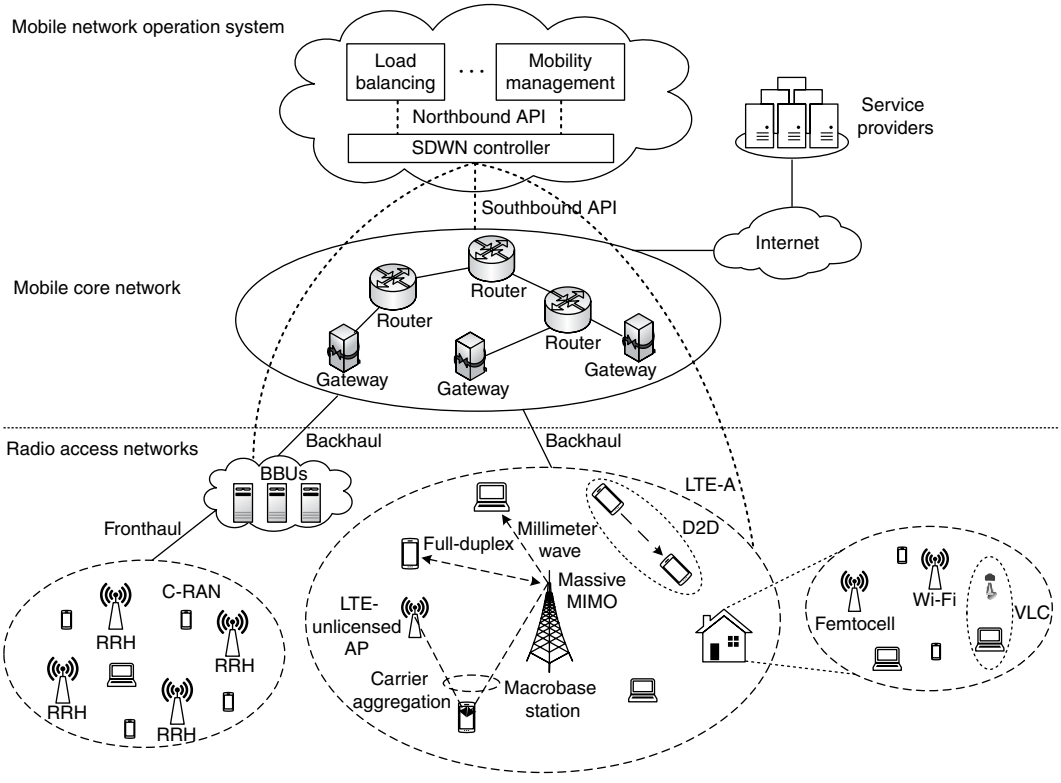

Figure 1.1 Illustration of the 5G network architecture. The radio access network includes various technologies such as C-RAN, massive MIMO, full duplexing, mmWave, femtocells, Wi-Fi, D2D, and VLC. The mobile core network can be controlled by a software-defined wireless network (SDWN) controller.

energy efficiency of 5G networks. In addition, different access technologies, including LTE and Wireless Fidelity (Wi-Fi), may be integrated to guarantee seamless coverage, and to support high data-rate transmission and data offloading.

In this chapter, we provide an overview of some of the exciting new technologies which are expected to be incorporated into 5G systems. These techniques will then be covered in detail in the subsequent chapters.

1.2 Cloud Radio Access Networks

It is reported [8] that peak traffic demand can be 10 times higher than off-peak traffic demand. However, as network resources for base stations are always provisioned for peak traffic demand, many base stations are lightly loaded or even in idle mode during off-peak hours, which leads to low utilization of the deployed cell sites. On the other hand, the energy efficiency of the lightly loaded base stations may also be low since the circuit power consumption constitutes a significant part of the total power consumption of a base station. C-RAN has recently been identified as a leading candidate for the 5G network architecture. In C-RAN, the baseband signal processing and the radio functionalities are decoupled. In general, a C-RAN consists of a baseband unit (BBU) pool placed in a cloud-based data center, and a large number of low-cost remote radio heads (RRHs) each deployed in a small cell. The BBUs and RRHs are connected through fronthaul links. The BBUs perform centralized signal processing and interference management. The RRHs retain only the radio functionality and communicate with the users over radio channels.

The C-RAN architecture has several advantages. First, C-RAN can adapt to spatial and temporal traffic fluctuations to provide on-demand services by exploiting the statistical multiplexing gain [9]. To this end, the number of BBUs required to support peak traffic demand can be reduced and the idle RRHs can be switched off to reduce power consumption, which leads to lower network capital expenditure (CAPEX) and operating expenditure (OPEX), respectively. Second, C-RAN facilitates the implementation of cooperative transmission/reception strategies, for example, enhanced intercell interference coordination (eICIC) and coordinated multipoint (CoMP) transmission [10]. With these cooperative strategies, spectrum efficiency can be significantly boosted via effective interference management among multiple RRHs. Third, C-RAN simplifies upgrading and maintenance of the network. Specifically, the virtualization of baseband signal processing on general-purpose processors or cloud servers simplifies network upgrades.

Despite the aforementioned advantages, C-RAN also poses new research challenges. First, in practice the fronthaul links have finite capacity, which can significantly degrade the performance gain achieved by C-RAN [11]. Second, to facilitate the centralized signal-processing and cooperative transmission strategies, massive amounts of accurate channel state information (CSI) are required at the BBUs [12]. In addition, user mobility leads to time-varying channels, which increases the CSI update frequency. Furthermore, owing to limited training resources and the transmission delay introduced

by the fronthaul links, the CSI received by the BBUs may not be accurate, which may degrade the ability to perform effective interference management. The C-RAN architecture will be discussed in detail in Chapter 2. An information-theoretic approach to determine the achievable rates of C-RAN with fronthaul capacity constraints will be presented in Chapter 3.

1.3 Cloud Computing and Fog Computing

Ubiquitous and pervasive computing services are crucial to the processing and storing of the significant amounts of data generated in IoT systems. The limited processing capacity of IoT objects may not always provide the computational power required for IoT applications. In this case, cloud computing can provide the necessary storage and processing capabilities. The cloud servers can collect data from different IoT devices, store the data, and run software applications to process and analyze the data. Cloud platforms such as ThingWorx, OpenIoT, Google Cloud, and the Amazon Web Services (AWS) IoT platform provide computing services for IoT application developers and service providers. Finally, the IoT service providers offer a set of services to IoT end users based on the information collected from IoT objects.

For delay-sensitive IoT applications with stringent latency requirements, conventional cloud services may not be appropriate. Fog computing [13–15], which is also known as *mobile edge computing*, extends cloud computing to the edge of the network. Here, IoT devices with processing and storage capability, called fog nodes, are deployed in the system and run applications on behalf of other devices. Since fog nodes are located in close proximity, the delay performance of computing services will be improved. In addition, fog aggregation nodes, which are network edge devices (e.g., routers and smart gateways) and have computing and storage capabilities, can provide further computing services to tasks that have more relaxed latency requirements. Figure 1.2 illustrates the coexistence of fog computing and cloud computing in IoT systems. In fact, fog computing is not a substitute for cloud computing. These two computing paradigms complement each other and together can provide the computational services required for IoT and improve the scalability of IoT applications. Mobile edge computing and fog computing will be discussed in detail in Chapter 4.

1.4 Non-orthogonal Multiple Access

Wireless communication systems have to provide services to many users in the same area (e.g., the same cell) concurrently. To coordinate and guarantee services for multiple users, some form of multiple access technique is required. The first four generations of cellular systems have relied on orthogonal multiple access (OMA). In particular, the first generation (1G) systems employed frequency division multiple access (FDMA). The second generation (2G) systems, which implemented the Global System for Mobile Communications (GSM), primarily used time division multiple

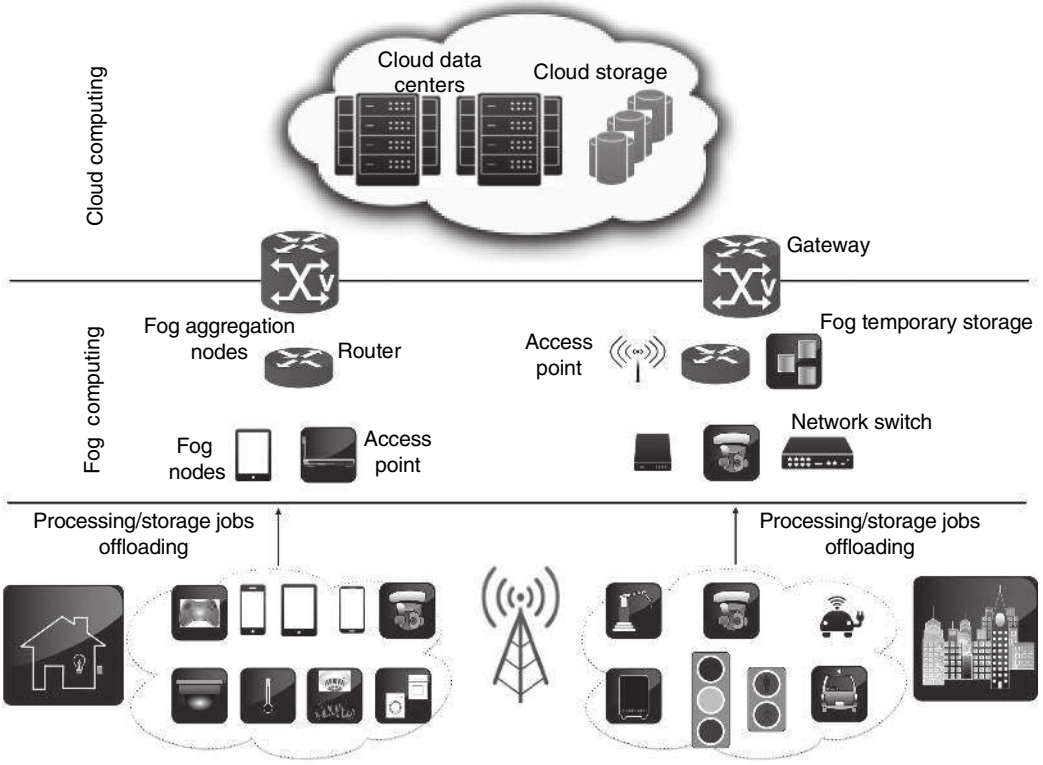

Figure 1.2 Fog computing and cloud computing can together provide computational resources for IoT objects and enable different IoT applications such as smart homes and smart cities. Fog nodes are located in close proximity to IoT objects and can respond to an emerging processing job quickly. Fog aggregation nodes, typically with higher processing power, can also support the computational requirements of IoT objects.

access (TDMA). The third generation (3G) systems, which implemented the Universal Mobile Telecommunications System (UMTS), relied on code division multiple access (CDMA). The fourth generation (4G) systems, which implemented LTE, adopted orthogonal frequency division multiple access (OFDMA). The main advantage of OMA is that under ideal conditions interuser interference is avoided, which significantly simplifies system and protocol design, including detection, channel estimation, and resource allocation. However, on the negative side, the number of users that can be supported in an OMA system is limited by the number of available orthogonal dimensions and, in practice, the orthogonality is often lost owing to effects such as frequency selectivity of the channel (in TDMA and CDMA) and phase noise and frequency offsets (in FDMA and OFDMA). Furthermore, from an information theory point of view, OMA is not optimal [16].

The shortcomings of OMA can be overcome by non-orthogonal multiple access (NOMA) techniques [17–21]. In NOMA, multiple users are scheduled on the same resource, i.e., in the same temporal, spectral, and spatial dimension. Thereby, a certain

amount of interuser interference introduced by non-orthogonal transmission is tolerated and removed at the receiver via successive interference cancellation (SIC). Because of its more efficient use of resources, both academia and industry see NOMA as one of the key enabling technologies for 5G systems [17, 19]. Although there are several different forms of NOMA, the two schemes that have received the most attention so far are power domain NOMA [18] and sparse code multiple access (SCMA) [17, 22].

In power domain NOMA, multiple users are multiplexed on the same time and frequency resource and power differences are exploited at the receiver for separation of the users' signals via SIC [20]. The advantages of power domain NOMA compared with OMA in terms of achievable data rate, coverage, and reliability were demonstrated in [21]. In addition, the combination of power domain NOMA with other 5G candidate technologies such as (massive) MIMO [23, 24] and FD transmission has been investigated. Although primarily a candidate for 5G, power domain NOMA has also been considered for standardization in 3GPP for downlink LTE transmission [19].

SCMA enables non-orthogonal access by overloading the multiuser system. Thereby, at the transmitter, the coded bits of a user are mapped to a complex codeword and the codewords of different users are overlaid using sparse spreading. At the receiver, joint multiuser detection and channel decoding using a message-passing algorithm is employed, where the sparsity of the spreading code limits the computational complexity [22]. Because of the overloading, SCMA can accommodate more users than OFDMA and achieve a higher throughput and connectivity.

NOMA is currently an active area of research and many challenges have not been fully resolved yet; examples include the fundamental information-theoretical limits of NOMA, channel coding and modulation design for NOMA, the integration of NOMA and other 5G techniques, including (massive) MIMO and full-duplex, security provisioning for NOMA, resource allocation for NOMA, and hardware implementation of NOMA. A thorough introduction to NOMA is provided in Chapter 6.

1.5 Flexible Physical Layer Design

As mentioned in Section 1.1, 5G networks aim to support not only voice and mobile Internet applications but also applications such as the IoT, M2M, the Tactile Internet, and vehicular communications. These applications have different requirements in terms of delay, reliability, and power consumption. In the last decade, orthogonal frequency division multiplexing (OFDM) and OFDMA have become the dominant physical layer technologies for providing high-data-rate services for major wideband wireless communication systems such as IEEE 802.11-based wireless local area networks (WLANs), Worldwide Interoperability for Microwave Access (WiMAX), and LTE-A. However, OFDM/OFDMA systems suffer from the following drawbacks: (i) OFDM-based systems have a high peak-to-average power ratio (PAPR), which requires the use of power-inefficient linear power amplifiers at the transmitter; (ii) the rectangular pulse shape of the OFDM symbol leads to large spectral side lobes and high out-of-band radiation; and (iii) subcarrier orthogonality is sensitive to carrier

frequency offsets and phase noise. Therefore, the existing OFDM technology may not be well suited for the transmission of the data of some 5G applications. Hence, several alternative non-orthogonal waveforms have been proposed and will be considered for the 5G physical layer. Promising candidates are filterbank multicarrier (FBMC), universal filtered multicarrier (UFMC), and generalized frequency division multiplexing (GFDM) [25]. These non-orthogonal signaling schemes attempt to overcome the limitations of OFDM/OFDMA by introducing new features into the signal and frame structure. For instance, GFDM is based on the modulation of independent blocks, and each block consists of a number of subcarriers and subsymbols. In GFDM, cyclic prefixes (CPs) and circular filtering are employed. In particular, GFDM exploits the "tail-biting" technique through circular filtering to decrease the length of the signal pulse tails. The circulant signal structure of GFDM also enables the use of one CP for an entire data block containing multiple GFDM symbols, which improves the spectral efficiency compared with conventional OFDM. In fact, GFDM is a flexible physical layer scheme since it covers both CP-OFDM and single-carrier frequency domain equalization (SC-DFE) as special cases. Besides, GFDM allows the time and frequency spacing of each data symbol to be adapted according to the channel properties and the type of application. The details of GFDM, including the receiver design and hardware implementation, will be presented in Chapter 7.

1.6 Massive MIMO

Multiple-antenna technology is a key element of current and future wireless communication systems. Traditionally, the number of antennas envisioned for MIMO communication systems has been limited to a comparatively small number, say 20 or less. However, recently it has been shown that multiuser MIMO communication systems exhibit several favorable properties if the number of antennas at the base station is increased to hundreds or even thousands [26], giving rise to so-called large-scale or massive MIMO systems. As a result, a large amount of theoretical and experimental research has been dedicated to massive MIMO communication systems since 2010.

It is well known that MIMO communication systems achieve substantial gains in spectral, power, and energy efficiency compared with conventional single-input single-output (SISO) systems. In fact, it has been shown that under ideal conditions the capacity of a point-to-point MIMO system with N_{T} transmit antennas and N_{R} receive antennas scales linearly with $\min\{N_{\mathrm{T}}, N_{\mathrm{R}}\}$, which is referred to as the multiplexing gain in the literature [27, 28]. However, point-to-point MIMO systems have several disadvantages in practice. First, the number of antennas that a mobile terminal (e.g., a smartphone) can accommodate is limited owing to size, power consumption, and cost constraints, which negatively impacts the achievable multiplexing gain. Second, the multiplexing gain may disappear altogether in the case of strong interference (e.g., at the cell edges), unfavorable channel conditions (e.g., insufficient scattering), and narrow antenna spacing mandated by the size constraints of mobile terminals. The disadvantages of point-to-point MIMO systems can be overcome by multiuser MIMO

systems [29–31]. In multiuser MIMO systems, a central node with multiple antennas (e.g., a base station) serves a number of (mobile) users with a small number of antennas. Thus, the signal-processing complexity at the mobile terminals is low, especially in the case of single-antenna terminals. In addition, since the users are spatially distributed over an entire cell, the angular separation of the terminals typically exceeds the Rayleigh resolution of the array and the channels of different users can be assumed independent. However, the multiple users in the system introduce interuser interference, which has to be mitigated by appropriate processing at the transmitter and receiver for downlink (i.e., base station to users) and uplink (i.e., users to base station) transmission, respectively. The uplink channel may be classified as a classical multiple access channel, for which many suitable linear and nonlinear receiver processing techniques are known from the rich literature on CDMA systems [32]. Thereby, while being computationally more complex, nonlinear receiver structures achieve a higher performance than linear ones. The downlink channel is a broadcast channel and suitable precoding techniques at the transmitter are needed to achieve high performance. Dirty paper coding was shown to be the optimal capacity-achieving precoding technique for a Gaussian MIMO broadcast channel [29, 30]. However, it entails a high computational complexity in practical implementation. Thus, linear precoding techniques such as zero-forcing (ZF) precoding, minimum mean-square error (MMSE) precoding, and regularized ZF precoding have attracted considerable attention as a good compromise between performance and complexity [33, 34].

For both uplink and downlink transmission, the availability of CSI at the base station is crucial for exploiting the full potential of multiuser MIMO systems. This is not critical for the uplink, where the users can simply send pilot symbols along with their data packets. In the downlink, channel estimation is more challenging and the best approach depends on the type of duplexing used. For frequency division duplex (FDD) systems, where different carrier frequencies are used for uplink and downlink transmission, the uplink and downlink channels are mutually statistically independent. Thus, each base station antenna has to first transmit pilots, which enable the users to estimate their respective downlink channels. Subsequently, each user has to feed back its channel estimate to the base station. Thus, assuming K users, the required number of feedback symbols grows linearly with $N_{\mathrm{T}}K$. On the other hand, for time division duplex (TDD) systems, the same carrier frequency is used for uplink and downlink. Thus, assuming a sufficiently large coherence time, the uplink and downlink channels are reciprocal and the base station can obtain the downlink channel by estimating the uplink channel based on pilots transmitted by the users. In this case, the number of pilots required grows linearly with the number of users K but is independent of the number of base station antennas.

Unlike conventional multiuser MIMO systems, which employ a comparatively small number of base station antennas (e.g., fewer than 20), massive MIMO systems are expected to employ hundreds or even thousands of base station antennas. Although such a tremendous increase in the number of antennas introduces new challenges in transceiver design and implementation, it has some interesting advantages for signal processing and communication. For example, if the number of base station

antennas is much larger than the number of users in the system, simple matched-filter (MF) precoding (downlink) and MF detection (uplink) at the base station lead to close-to-optimal performance, facilitating low-complexity signal processing at both the base station and the user terminals. Nevertheless, as will be shown in Chapter 8, as the number of users increases, significant performance gains can be achieved with ZF and MMSE precoding and detection schemes. Furthermore, random impairments such as small-scale fading and noise are averaged out as the number of base station antennas grows large. To keep the signaling overhead for CSI acquisition in massive MIMO systems manageable, TDD operation is preferred, since for FDD systems the amount of CSI feedback grows with the number of base station antennas. However, a major impairment in massive MIMO systems is so-called *pilot contamination*. Pilot contamination is caused by the reuse of the same (or linearly dependent) pilot sequences in different cells. This reuse is unavoidable as, for a given pilot sequence length, the number of linearly independent pilot sequences is limited. Recently, several efficient techniques have been proposed to overcome pilot contamination [35, 36].

Massive MIMO also has the potential to significantly improve energy efficiency. It has been shown [37] that if N_{T} grows large and all other system parameters are assumed constant, the transmit power per user in multiuser massive MIMO systems can be reduced proportionally to $1/N_{\mathrm{T}}$ and $1/\sqrt{N_{\mathrm{T}}}$ for perfect and imperfect CSI knowledge respectively, at the base station, without affecting throughput and reliability. Hence, massive MIMO systems offer a simple path to more energy-efficient and "greener" communication networks. Furthermore, a major concern for future wireless communication systems is security and privacy. Massive MIMO is well suited to addressing these issues. In fact, because of the large number of spatial degrees of freedom, massive MIMO can be exploited to protect cellular systems against passive [38] and active [39] eavesdropping.

Because of its favorable properties, massive MIMO is expected to be one of the core technologies of 5G systems [3]. Nevertheless, massive MIMO still has many challenging open research problems. For example, because of the large scale of massive MIMO systems, the use of cheap hardware components is desirable. However, this in turn gives rise to hardware impairments, such as phase noise, in-phase/quadrature phase imbalance, and amplifier nonlinearities, which have to be properly dealt with to avoid performance degradation [40]. Furthermore, the channel-hardening effect induced by the large number of base station antennas necessitates the design of new resource allocation and user association algorithms. Massive MIMO systems and scheduling protocols will be discussed in detail in Chapters 8 and 15, respectively.

1.7 Full-Duplex Communications

Although advanced technologies such as C-RAN, NOMA, and massive MIMO are able to alleviate system resource shortages, the spectral resource is still underutilized. In particular, traditional wireless communication devices operate in the half-duplex (HD) mode, where downlink and uplink communications are separated orthogonally

in either frequency or time, which results in a significant loss in spectral efficiency. Although researchers have proposed various techniques for minimizing/recovering the spectral-efficiency loss inherent in HD communication, such as joint dynamic uplink and downlink resource allocation [41] and two-way HD relaying [42], these schemes do not solve the problem fundamentally, since the associated protocols still operate in the HD communication mode.

Recently, full-duplex (FD) wireless communication has emerged as a candidate technique for 5G networks and has received significant attention from both industry [43, 44] and academia [45–47]. Compared with existing communication networks adopting HD transmission, FD systems simultaneously transmit and receive data signals in the same frequency band, which has the following advantages [45]. First, FD systems can provide a better utilization of time and frequency resources such that it is possible for FD systems to double the link capacity compared with existing HD systems. Second, in practice, in addition to data signals, feedback signals such as control information or CSI can be also transmitted concurrently to facilitate data communication. Thus, FD systems can reduce the feedback delay by receiving feedback signals during data transmission. Third, FD systems can improve communication security. In fact, an HD base station is unable to guarantee physical layer security in the uplink unless external helpers perform cooperative jamming to interfere with potential eavesdroppers. In contrast, an FD base station can guarantee secure uplink transmission by transmitting jamming signals in the downlink while receiving the desired uplink information signals. Last but not least, a hybrid HD and FD protocol, which retains the option to utilize one frequency band for FD communication or two orthogonal frequency bands for two parallel HD communications, can be adopted to increase the flexibility in spectrum usage.

Despite the potential benefits, the performance of FD communication systems is limited by self-interference (SI), which is caused by signal leakage from the downlink transmission to the uplink signal reception (Figure 1.3). In particular, the ratio between the SI power and the desired incoming signal power can easily exceed 100 dB [48]. The huge difference between the signal powers causes saturation in the analog-to-digital converter (ADC) at the receiver front end of FD devices which severely jeopardizes signal reception. Hence, FD communication has been considered impractical for the past 60 years. Fortunately, recent research has shown that FD communication is feasible by using spatial SI suppression, digital/radio frequency (RF) interference cancellation techniques, and transmit/receive antenna isolation [47]. Several prototypes of FD transceivers using various SI cancellation techniques have been built to demonstrate the feasibility of FD communication and the expected performance gains compared with HD communication in different physical environments [49–51].

In fact, FD technology introduces new research challenges for wireless communication engineers in both resource allocation and communication protocol design. In the following, we briefly discuss some open issues in FD communication systems. In general, strong SI is an obstacle in realizing FD communications since it increases with the transmit power of the FD devices. As a result, multiple-antenna technology has been proposed to overcome SI. In particular, by utilizing the extra spatial degrees of