1

Introduction to 5G and Beyond Network

We have witnessed an unprecedented development of wireless technology for the past few decades. Starting from 1980s, when the first mobile phone was released, major wireless technology advanced almost every decade. From first generation (1G) to 4G. The invention of smart devices, such as phones, tablets, and home appliances, is the main driving force for the ever-increasing mobile traffic today. It is not surprising that mobile traffic increased 10-fold between 2014 and 2019 globally. The mobile data traffic is expected to grow much faster than fixed IP traffic in the upcoming years [34]. Wireless technologies dramatically changed the way people interact, communicate, and collaborate, especially at post-Covid era. The need for faster, more efficient and secure, and intelligent communication technique remains strong. While the current wireless communication systems such as 4G long term evolution (LTE) have been pushed to their theoretic capacity limit, different air interface and radio access technologies including heterogeneous network (HetNet) [76, 77], multiuser multi-input multi-output (MU-MIMO) [105], and device-to-device (D2D) communication [51] have become potential paradigms to fulfill the gap between demands from end users and the capacity that current air interface can provide.

1.1 5G and Beyond System Requirements

In their pioneering work [10], Andrews *et al*. evaluated the requirements for 5G. In short, 5G wireless communication system should provide 1,000 times aggregate data improvement over 4G, support for as low as 1 ms round-trip latencies, 10 times longer battery life for low-power devices, and also support 10,000 times or more low-rate devices in a single macro cell, see Figure 1.1 for a brief illustration. Due to those high requirements, the transformation from 4G to 5G cannot be simply fulfilled by extensions of current technologies. In general, 5G and beyond system should support or deliver the following aspects. Notably, (i) more bandwidth. Currently commercial cellular systems use frequencies below 6 GHz (sub-6 GHz); in fact, there is abundant bandwidth in the millimeter-wave (mmWave) band, for example in 28 GHz and above, which can provide more bandwidth that previously have not been applied in cellular networks. (ii) More antennas. Higher frequency also brings smaller form factor of large antenna arrays. Additionally, the signal processing techniques in terms of massive MIMO and transceiver design also improved

5G and Beyond Wireless Communication Networks, First Edition. Haijian Sun, Rose Qingyang Hu, and Yi Qian.

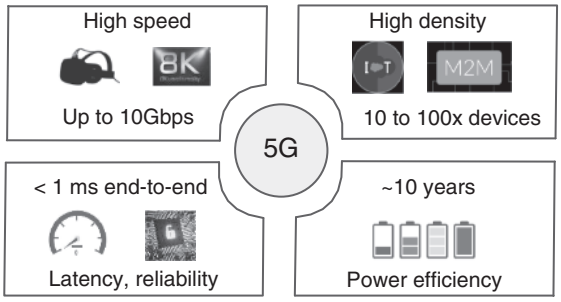

Figure 1.1 Four main goals for 5G.

significantly. (iii) New radios (NR). The physical layer in 5G will change dramatically, specifically the 5G NR, which includes the new multiple access technology, the new air interface, and a combination of several existing techniques. (iv) New schemes. It is expected that ultra dense networks (UDN) will be heavily deployed. The density of small base station (BS), such as micro BS, femto cell, and pico cells, will be much higher than that in 4G. But they share the similarity in terms of deploying BSs with different powers to provide seamless coverage, as well as performance improvements from short-range communications. (v) High intelligence. It is expected that beyond 5G systems should support higher level of intelligence. Emerging applications such as Artificial intelligence (AI), semantic communication, and robots will surely benefit from AI-friendly wireless technology. (vi) Pervasive wireless. It is anticipated that each person will carry more personal devices for enhanced life style and health monitoring. To support ubiquitous wireless connectivity, those devices need be connected. Current network architecture can hardly support such high number of devices simultaneously.

1.1.1 Technical Challenges

The above promising technologies are able to deliver ambitious goals of 5G, but they ultimately encounter some challenges. First of all, even though high-frequency bands have major vacancy, mmWave signals are notorious for weak penetration and vulnerable blockage; hence, the transmission characteristics are big concerns. Moreover, studies also have shown mmWave signals have high attenuation due to atmospheric gaseous, rain, concrete structure, glasses, even foliage. The real-world deployment of such mmWave systems needs to be carefully studied and planned. Secondly, from the transceiver design perspective, higher-frequency signals impose challenges in circuit design, materials, and heating issues. Nyquist theorem sets the lower boundary for sampling rate in communication systems. With wide bandwidth in mmWave spectrum, sampling rate can reach up to 10 Gbit/s level, and high-speed circuit design becomes very difficult. It is also reported that the energy efficiency for components (power amplifier, analog-to-digital converter, digital-to-analog converter) in high frequency is low, only around 10%. One of the major concerns from network operators is that power consumption will hike due to 5G. Furthermore, the low efficiency in these components also brings thermal issues in hand-held devices, degrading user experiences. Thirdly, with mmWave band, performance gain largely comes from large-scale antenna array, current design can integrate hundreds of antenna elements in a small area (due to small wavelength of mmWave signals). Even though this can facilitate

the beamforming, which generates narrow but stronger signals toward desired direction, the overhead for channel estimation, precoding, and beam tracking is too large. Fourthly, in UDN networks, since the transmitter density is high, signals can cause higher interferences with each other. The problem will be more severe with high-density users in the same area. Challenges in mobility management, interference management, and heterogeneity nature of devices are severe. Lastly, it is expected to support intelligent applications in beyond 5G systems. For example, conventional communication systems are transparent of message (i.e. they are only responsible for transmitting bits but do not know any further info). Semantic communication, on the other hand, has knowledge of the underlying message, and the communication scheme can be dynamically changed to fit different needs of the message. Besides, ubiquitous wireless signals open door for sensing applications, such as localization, monitoring, and healthcare. In recent years, intelligent communication system has been proposed to accommodate these needs. A notable example is wireless federated learning system to cater the distributed machine learning. However, a deep integration from wireless design perspective is strongly desired.

Recently, there are several emerging technologies which aim to deliver the goal of 5G and beyond, and address the challenges above. Specifically, in this book, our focus is on the physical layer techniques, such as 5G NR non-orthogonal multiple access (NOMA) and physical layer (PHY) mobile edge computing (MEC), high-level communication architecture for pervasive Internet of Things (IoT) devices, as well as wireless federated learning system design. We have conducted preliminary researches to address the challenges mentioned above. Specifically, we discuss how to utilize NOMA on improving aggregated data rate and supporting more devices simultaneously, propose schemes for wearable IoT communications, discuss the usage of MEC on helping with power consumption and latency, and analyze how wireless design can facilitate distributed machine learning. Below we briefly introduce each enabling technique.

1.2 Enabling Technologies

1.2.1 5G New Radio

1.2.1.1 Non-orthogonal Multiple Access (NOMA)

Initially proposed by NTT DOCOMO as an enhancement for LTE-advanced (LTE-A) in 2013, NOMA has been recognized as one of the most promising techniques for 5G due to its capability of supporting a higher spectral efficiency (SE) and native integration of massive connectivity. The basic principle of NOMA is that at the transmitter side, multiple signals are added up with different powers, forming a superimposed signal (SS). To ensure weak user's quality of service (QoS), at the receiver side, successive interference cancellation (SIC) is used to retrieve each user's signal sequentially from the SS. Specifically, a user can decode the strongest signal by treating other signals as interference. If the decoded signal is its own data, SIC stops. Otherwise, the receiver subtracts the decoded signal from the SS and continues to decode the next strongest signal. Notice that SS with SIC is not new; in information theory, this duo is a capacity-achieving technique in the uplink communication. However, the difference is in NOMA, the weak user has a stronger power, which

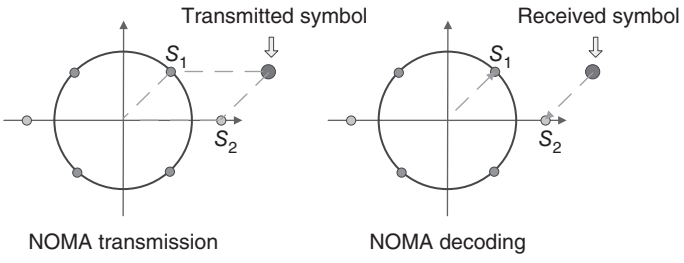

Figure 1.2 NOMA principles: transmission and decoding stage.

is not information-theoretic optimal. Since its design philosophy may be combined with diverse transceivers, it has drawn tremendous attention in multiple-antenna systems and in downlink and uplink multi-cell networks. In contrast to classic orthogonal multiple access (OMA), NOMA provides simultaneous access to multiple users at the same time and on the same frequency band, for example by using power-domain multiplexing. It has been shown that NOMA is capable of achieving a higher SE and energy efficiency (EE) than OMA, such as OFDMA, time division multiple access (TDMA), and frequency domain multiple access (FDMA). Figure 1.2 shows the basic principle of NOMA with data encoding and decoding. S_1 and S_2 are the symbols for users 1 and 2, respectively. We also assume user 1 has a better channel than user 2. At the transmitter side, binary phase shift keying (BPSK) and quadratic phase shift keying (QPSK) modulation are applied, respectively, for the two users. Clearly, the average symbol power of S_2 is larger to compensate for the unfavorable channel. Actual transmitted symbol is simply the addition of these two. At the receiver side, symbols with the highest power will be decoded first, in this example, S_2. Besides, since the received symbol is on the right side of y-axis, for BPSK, it will be decoded as S_2, and then removed from the composite signal, which only has S_1 left. Notice that the symbols can use the same modulation scheme as long as they have different power. Most NOMA works, however, do not consider any specific modulation, rather they apply the Gaussian coding and perform analysis based on information-theoretic perspective.

The disadvantage of NOMA, however, lies in the following aspects. Firstly, NOMA requires a more complicated receiver structure to perform SIC; hence, the cost will be higher and receiver architecture will also be changed accordingly. Secondly, during SIC procedure, one user will decode signal from others; this will cause security and privacy concerns. Lastly, depending on implementation, this successive decoding will impose certain delays for users.

Starting from 3rd Generation Partnership Project (3GPP) LTE Release-13, NOMA, as one of the techniques in multi-user superposed transmission (MUST), has become part of the standardization. In 2017, with LTE Release-14, 15 MUST schemes have been proposed for the uplink NR. Additionally, NOMA has attracted extensive attention from industry. NTT DoCoMo and MediaTek collaborated to have a field test of NOMA in Nov. 2017. With a simple scenario of one base station and three users, they were able to achieve 2.3 time spectral efficiency compared with current technology.[1]

1 MediaTek Newsroom, Nov. 2017.

Nevertheless, we have applied NOMA in many schemes and systematically studied its performance, for example NOMA with D2D, with MIMO, relay networks, and cognitive radio. More importantly, we have reviewed the fundamental principle of NOMA and pointed out the error propagation phenomenon. Furthermore, we have also considered the channel imperfection and its impact to NOMA performance.

1.2.1.2 Channel Codes

Channel coding is instrumental for achieving higher capacity and reliability. For example, low-density parity-check (LDPC) has been extensively used in 4G, replacing convolutional and turbo codes in previous generations. In 5G NR, polar codes are identified as another promising capacity-achieving coding technique. Polar codes have been adopted in 5G standardization process. For example, 3GPP incorporates polar codes for both uplink and downlink control information communication service, such as enhanced mobile broadband (eMBB), massive machine type communications (mMTC), and ultra-reliable and low latency communications (URLLC). Channel codes for 5G NR should be flexible to support the variable rate and length for both data and control packets. To address that, LDPC has developed several variations, such as quasi-cyclic (QC) LDPC codes for better rate matching and interleaving, as well as parallelism for efficient encoding and decoding [59]; Multi-edge (ME) LDPC mainly for throughput improvement and can scale well in larger block lengths. On the other hand, newly introduced polar code takes advantage of channel polarization, a natural behavior due to signal propagation. Correspondingly, encoding is recursively performed by the channel transformation matrix and creates channels that are either perfectly noiseless or completely noisy. A detailed tutorial of polar codes can be found in [16].

1.2.1.3 Massive MIMO

Massive MIMO refers to applying large-scale antenna elements at transmitter and/or receiver side, usually the number of antenna is hundreds or more. MIMO can exploit spatial diversity or multiplexing, and improve system reliability (for example, lower bit error rate) and throughput, respectively. Compared with legacy MIMO system, massive MIMO brings significant improvements in diversity and multiplexing to fully exploit wireless channel characteristics. One prominent aspect is massive MIMO can generate very narrow beams toward the receiver side. Hence, it can not only increase reception power, but also benefit network capacity and coverage, and ultimately provide better user experience.

These benefits come at a price. Like MIMO, performance gain from massive MIMO largely comes from beamforming and advanced signal processing techniques, which require channel information. If both transmitter and receiver have massive MIMO antennas, their channel is a matrix with hundreds by hundreds of elements. Overhead for accurate channel estimation is prohibitively large. For example, orthogonal pilots are usually applied to obtain channel information; in the case of massive MIMO, maintaining pilot orthogonality is difficult, not to mention practical challenges such as pilot contamination and offset (time and frequency). To address these challenges, prior works have studied robust beamforming design, such that the requirement for accurate channel information can be relaxed. Furthermore, signal processing in massive MIMO is also sophisticated. Traditional optimization methods for throughput maximization or bit error rate (BER)

minimization become problematic due to high computation complexity, which hinders the deployment in mobile devices.

It is worth to note that other approaches such as applying out-of-band information, including vision, location, and geometry data to assist beamforming are also studied. Out-of-band information provides complementary details for assisting beamforming steering. These emerging solutions are primarily motivated and enabled by machine learning.

1.2.1.4 Other 5G NR Techniques

5G NR also has other innovations. Recent 3GPP releases 15, 16, and 17 gradually bring more flexibility and enhancement on several aspects. For example, dynamic slot structure caters to different communication needs, for either low-latency or high data-rate application. This structure allows for customized slot design, for examples, adding a longer or shorter cyclic prefix, changing the data frame length, or providing extra guard space. Another innovation is spectrum sharing. In contrast to static database-aided spectrum sharing, which detects secondary users' interference and only allows them to access bands in an opportunistic way, current spectrum sharing is more dynamic, enabled by advanced machine learning-based approach, hence is more efficient and accurate.

1.2.2 Mobile Edge Computing (MEC)

Due to the size, battery, and cost limitations, mobile devices can experience performance bottleneck when computation-intensive tasks are added. More than one decade ago, people solved this problem by introducing the concept of cloud computing. Mobile devices do not perform large-scale computation locally; instead, they send these tasks to remote servers for faster and more secure processing, storage, and sharing. The centralized nature of cloud-based computing can reduce the expenditure cost while providing easier deployment process. However, cloud servers may be located in remote areas, which causes inevitably longer end-to-end transmission and processing delay.

MEC is a new alternative paradigm for the upcoming 5G systems. Instead of transmitting data to the remote servers for processing, MEC provides certain computation capacities locally, for example within the base station in the wireless cellular networks. This paradigm shift can effectively reduce long backhaul latency and energy consumption, as well as support a more flexible infrastructure in a cost-effective way. MEC has attracted extensive research interests recently, not only in the architectural level, but also in specific tasks such as cooperative computation offloading. Computation offloading, which leverages the powerful MEC servers in proximity and sends the computation-intensive tasks for further processing, is a desirable scheme to overcome the physical limitations of user devices (Figure 1.3).

We see this paradigm shift in a more fundamental way. In cloud computing era, even though the data transmission speed is not high, the bottleneck comes mainly from the computation capacity. With Moore's law still being effective, performance of integrated circuit chips grows exponentially. On the other hand, communication technology makes the speed increase almost linearly. Since the goal is to reduce processing speed, it is more beneficial to perform task execution both locally and remotely.

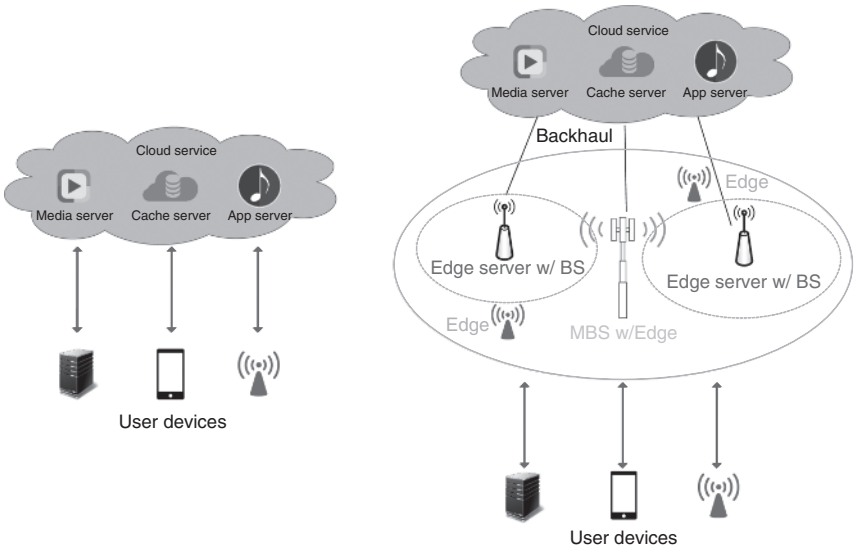

Figure 1.3 Paradigm shift from cloud computing to mobile edge computing.

In order to reduce latency as well as to improve system efficiency, we propose a joint processing scheme in which the total task can be divided into two parts, one for local computing and the other for offloading. To cope with the ever-increasing concerns on energy efficiency, we evaluate the system performance by a new metric, computational efficiency (CE). It is defined as the total number of bits computed with the total energy consumption. The objective is to maximize each user's CE with time constraints (users should finish their task before a required time), energy constraint (each user is powered by battery; hence, the total energy should be below a threshold), and task constraint (each user should finish a minimum number of data bits). Later we show CE is a more appropriate method in terms of finding the balance of more tasks and less energy.

1.2.3 Hybrid and Heterogeneous Communication Architecture for Pervasive IoTs

Recent years have witnessed the unprecedented growth of wearable devices owing to the swift advances in chip design, computing, sensing, and communications technologies. While wearable devices are not new, the past few years have seen a surge in their large-scale use and popularity. A wearable device or simply a wearable refers to a device that can be worn on the body. This rapid rise in popularity was spurred, in part, by technological innovation. Emerging system on chip (SoC) and system in package (SiP) have scaled down the printed circuit board (PCB) size, decreased power consumption, and most importantly, have made it possible to design wearables in a variety of desired shapes (Figure 1.4). Wearable devices provide easier access to information and convenience for their users. They have varying form factors, from low-end health and fitness trackers to high-end virtual reality (VR) devices, augmented reality (AR) helmets, and smart watches. These devices can collect data on heart rates, steps, locations, surrounding buildings, sleeping

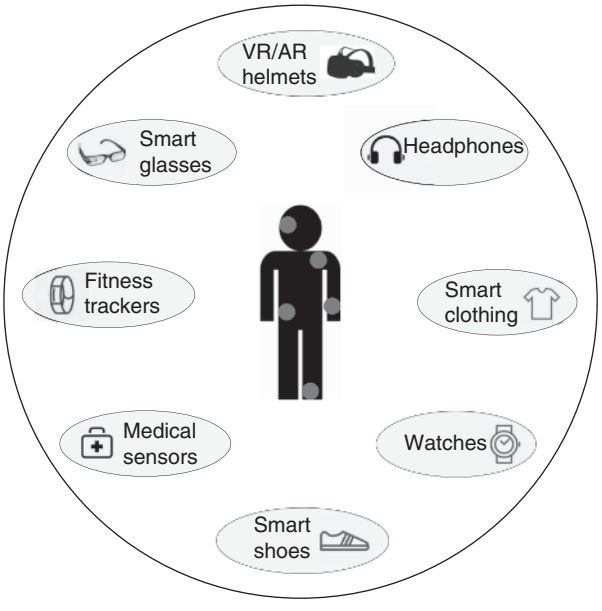

Figure 1.4 Wearable devices may have varying forms, from small medical sensors to entertainment helmets.

cycles, and even brain waves. Yet computing limitations continue to hinder wearables' ability to process data locally. As a result, most devices choose to offload their collected data to other powerful devices or to the clouds. This necessary communication plays a key role in wearable devices. Different applications provided by different wearables may have varying communication requirements. For example, while medical sensors have stringent requirements on latency and reliability, they have a low data rate need. On the other hand, AR/VR devices need both high throughput and low latency for a better user experience.

Wearable devices may not be able to take full advantage of current communication architecture, due to their potential cost and hardware complexity. On the other hand, wearable devices have succeeded in becoming more and more involved in everyday activities requiring voice, image, and video inputs. Human beings are generally sensitive to an approximate 100 ms audible delay and can catch visual delays of less than 10 ms. Furthermore, cell phones and tablets now use primarily touch interaction, a "tactile interaction" that requires a more rigorous delay control, such as 1 ms. A promising heterogeneous and hybrid network architecture is shown in Figure 1.5. It contains small BS (SBS), marco BS (MBS), remote radio head (RRH), and network slice.

1.3 Book Outline

In face of several challenges by 5G and beyond system, this book focuses on technologies that can improve spectral, energy, and computation efficiency. As mentioned above, we

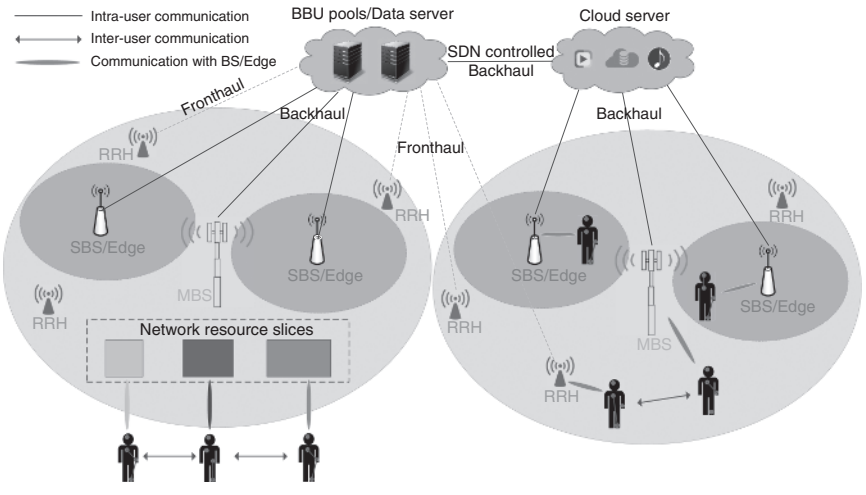

Figure 1.5 A promising network architecture for pervasive IoT communication needs.

mainly study physical layer techniques. Specifically, our first focus (Chapters 1–6) is to provide reader with latest research efforts on 5G NOMA. We have studied NOMA in a systematic way, including applying NOMA to address spectral efficiency and number of connected devices challenges under various network schemes. Our next focus (Chapters 7 and 8) is MEC. MEC is used to reduce computation delay, and we primarily investigate its role for computation offloading. Chapter 9 discusses the emerging wireless paradigm to facilitate distributed machine learning. Chapters 10 and 11 review secure spectrum sharing with machine learning techniques. Lastly, Chapter 12 concludes this book and discusses current and further research directions on 5G and beyond wireless systems.