1 Introduction

1.1 Why Green and Soft?

For the past forty years, mobile communication systems have been undergoing a revolutionary change from the first-generation (1G) analog cellular systems to the fourth-generation (4G) long-term evolution (LTE) systems. 4G could provide high-speed data service, including internet access, high-definition video broadcasts, and so on. With the development of mobile internet and service diversity, wireless traffic is growing rapidly, especially in developing countries, as shown in Fig. 1.1.

From Fig. 1.1, we can predict that the data explosion will continue in the future, driven by the vigorous development of mobile internet and internet of things (IoT). More and more mobile internet applications have emerged to meet the diverse demands of subscribers. The fifth-generation (5G) wireless networks will touch many aspects of our daily life in the future, such as home, work, leisure, and transportation. As a consequence, a consistent service experience should be supported in various scenarios, including dense residential areas, office buildings, stadiums, open-air gatherings, subways, highways, high-speed trains, and wide-area coverage scenarios. IoT is focused on communications between things and things, and between things and people, involving not only individual users, but also a large number of various vertical industrial customers. IoT applications are usually complex and diverse, therefore 5G should be more flexible and more scalable, to support massive device connections and meet diverse requirements.

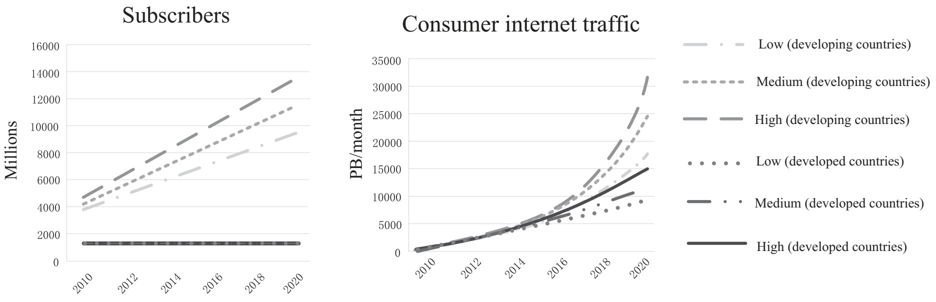

Figure 1.1 The growth of subscribers and consumer internet traffic.

The very challenging requirements from mobile traffic in these scenarios are characterized by ultra-high traffic volume density, mobility, or connection density, etc. The key performance indicators (KPIs), as defined by the International Telecommunications Union (ITU) [1], include peak data rates of 20 Gb/s, user experienced data rates of 100 Mb/s, support for up to 500 km/h mobility, 1 ms latency, a connection density of 10^6 devices/Km^2, a network energy efficiency improvement of 100X, and an area traffic capacity of 10 Mb/s/m^2. Though all the requirements need not be met simultaneously, the design of 5G networks and radio access should provide flexibility to more efficiently support various applications in diversified scenarios. In addition to the technical requirements on 5G, the operators are also faced with the requirement of taking social responsibilities when deploying the 5G networks. The first one is how to tackle the global warming issue. As it is becoming more and more serious, the impact of wireless communication networks on the environment has drawn extensive concerns. It has been reported that worldwide information and communication technology (ICT) contributes around 2 percent of the global carbon dioxide emissions (comparable to the global aviation industry), and it is estimated to grow to 4 percent by year 2020 [2]. In addition, ICT accounts for about 10 percent of global electricity consumption, and mobile network is one of the significant components of ICT energy consumption. Therefore, 5G mobile networks should be more energy-efficient than ever before to reduce both the operational costs and carbon dioxide emission. Motivated by this, network operators, regulatory bodies, such as the 3rd Generation Partnership Project (3GPP) and ITU, have conducted several research activities aiming at improving the network energy efficiency (EE). A lot of work related to green communication has been done, such as the Mobile Virtual Centre of Excellence (VCE) Green Radio project, EARTH project, and the international Greentouch Consortium [3, 4].

Another issue is environmental pollution from the outdated infrastructure equipment. The transition of mobile communication systems from one generation to another occurs generally at the expense of abandoning huge amounts of equipment in either core networks or radio access networks of the previous generations, which may pollute the environment if not handled properly. Is there a possibility that such generation transition can be conveniently and efficiently achieved via software upgrade, without abandoning old hardware or replacing it with newly manufactured equipment? The 5G network is therefore motivated to be reconfigurable with software-defined networking (SDN) [5] and air interface agility in implementation.

In the past several decades, high capacity and spectral efficiency (SE) are the primary design goals of mobile network, but now we need to pursue a SE–EE codesign network. Besides satisfying diverse user demands, future mobile communication systems should be able to support lower power consumption to build a greener mobile communication network and achieve greater sustainability. Meanwhile, the 5G network needs to facilitate a converged network synergistic with information and communication technology convergence, multiple radio access technology (RAT) convergence, radio access network (RAN) and core network convergence, content convergence, and spectrum convergence.

1.2 Green: From UE to Infrastructure

Reducing carbon emissions and operating expenditure (OPEX) costs are important goals for wireless cellular networks. The profound meaning of green is to heighten efficiency in utilization of any resources supporting wireless communications from the network side to the user terminal (UE) side.

For the UE side, the required energy in the UE's battery is increasing with the development of mobile internet. How to optimize the battery life of mobile users is still a challenging task. To solve this problem, several methods have been proposed. For example, power-saving mode (PSM), such as discontinuous reception (DRX) mechanism, has been introduced in LTE for power saving at the UE. DRX enables the UE to switch from an active state to a short or long sleep state without sacrificing the quality of service (QoS). In the sleep state, a terminal remains attached with the network. However, it is not accessible because it does not check for paging.

In 5G, a new user mode, called RRC_INACTIVE mode [6], is introduced. In the RRC_INACTIVE mode, the terminal can return to communication state without RRC connection setup procedure, and hence the energy consumption can be reduced further.

To cope with the limited battery energy problem, radio frequency (RF) energy-harvesting technology has garnered extensive attention recently [7–9]. Since the RF signals radiated by transmitter carry both information and power at the same time, it is natural to think that the devices can be powered by the energy from the received electromagnetic waves. An example of a simultaneous wireless information and power transfer (SWIPT) system is shown in Fig. 1.2. In this system, the UE intends to decode the information and harvest energy from the received signal simultaneously. The power of the received signal at the UE is split into two parts, for decoding the information and for energy harvesting. With the energy-harvesting mechanism, the UE is expected to be not only environmentally friendly but also self-sustainable.

Recently, many mobile application developers have also shown interest in how to prolong the battery life. For applications involved with heavy network usage, such as online video, an appropriate mobile network mechanism can be designed to avoid sustained mobile network signaling interaction, frequent small data transaction, and so

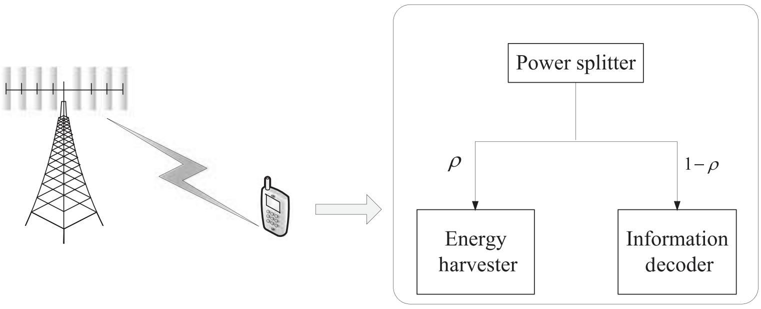

Figure 1.2 SWIPT system.

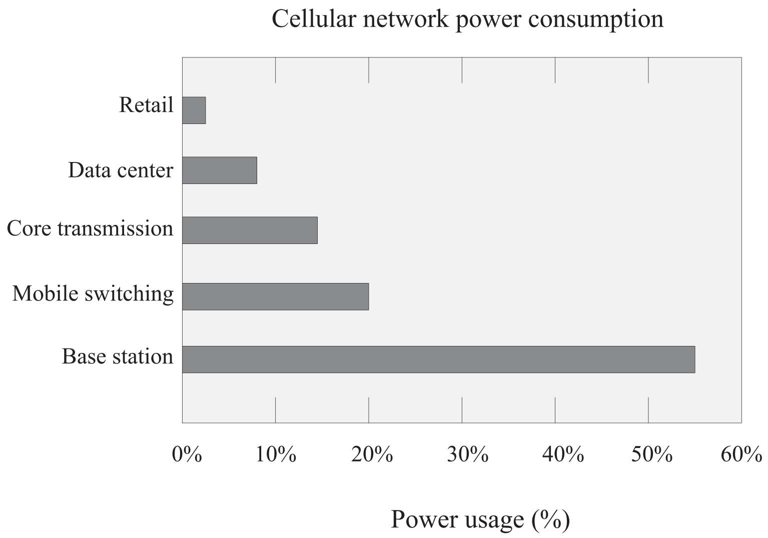

Figure 1.3 Cellular network power consumption.

on, and hence reduce power consumption. For applications requiring short interaction time and less data transaction, such as game play, the power consumption is mainly from an application processor (AP) module. Therefore, the optimization of an AP part should be considered.

As shown in Fig. 1.3 [10], based on the research of the Green Radio program, a base station (BS) takes up more than 50 percent of the cellular network power consumption. Therefore, most of the projects mainly focus on energy-saving issues at the BS and put a lot of effort into developing innovative techniques for reducing the energy to operate a RAN and to design appropriate radio architectures for energy saving.

To improve the energy efficiency of wireless networks, the Mobile VCE Green Radio Project [8] has proposed several techniques and concepts, including higher-efficiency antennas, power amplifiers (PA), multi-hop relaying techniques, BS cooperation, interference cancellation techniques, as well as energy-aware packet transmission and scheduling protocols. The results suggest that as much as 90 percent of the overall energy can be reduced under high-traffic conditions when combining high-efficiency antenna, PA, coordinated multipoint (CoMP) techniques, and shifting the network topology to heterogeneous networks (HetNets). The EARTH project [9] has investigated the energy-saving problem from both network aspects and radio components level. For the network level, new deployment strategies of RANs, self-organizing management mechanisms, and signaling protocols for energy efficiency optimization have been considered. For the component level, the power-efficient transceiver that can adapt to traffic load has been studied. For example, the supply voltages can be changed with estimated average signal power of incoming baseband signals, and some baseband boards can also be switched off. The GreenTouch project has also conducted a lot of studies to improve the energy efficiency from different aspects, including the mobile networks, the fixed access networks, and the core networks [11]. To make the mobile access networks more energy efficient, several schemes have been suggested. By decoupling the control and data plane functionalities, the small cells can be turned on and off based on the traffic load to save energy consumption. Antenna-sleeping technology is utilized in dynamic multiple-input and multiple-output (MIMO) systems to enable the BSs to switch between single-user and multiuser mode with the optimal number of active antennas. For fixed-access networks, a Gigabit Passive Optical Network (GPON)-based fiber-to-the-premises (FTTP) solution and redesigned low-power optical transceiver for optical access are brought together. To burst the energy efficiency of a core network, power-saving network components have been investigated, including energy-efficient routers, transponders, and improved digital signal processors. In addition, the relationship between the traffic demand and the power consumption of routers and transponders is used as guidance to find the optimal combination of routers and transponders. The study of GreenTouch has demonstrated that the net energy consumption in end-to-end communication networks can possibly be reduced by up to 98 percent by 2020 while taking into account the traffic growth between 2010 and 2020.

In the chip level, many new materials have been utilized for energy saving. In PA design, GaAs (gallium arsenide) transistors are usually used because of their ability to operate at high frequency, and they can generate signals with lesser noise. Currently, the GaN transistor has attracted great attention of manufacturers due to its superior characteristics of high output power, high breakdown voltage, and high temperature stability [12]. It has been reported that using GaN technology allows more than six times the output power of existing PA, using GaAs transistors. Besides, how to improve the cooling capacity without increasing energy consumption is crucial for both UE and network infrastructure. It has been proved that graphene-based materials have better thermal dissipation ability, therefore the graphite sheet has been widely used as thermal-averaging material in terminal design to delete the extra high temperature point.

1.3 Soft: From Core Network to RAN

Unlike software upgrade, it usually takes a long time to evolve to a new communication system, since the launch cycle of new standards is long, and new equipment needs to be developed and integrated. Therefore, it is necessary to make the network more flexible and reconfigurable. Soft design is the key to achieve these goals, and it will bring agility to the implementation of network elements from both core networks and access networks, as well as the building blocks of air interface. In a soft network, computing, storage, and radio resources are virtualized and centralized in order to reach dynamic and user-centric resource management.

The soft idea in communication networks can be traced back to the 1990s. In early 1990, the communication industry realized that it was difficult to define a unique standard for future mobile systems, and hence software-defined radio (SDR) emerged [13]. Some components of radio systems are implemented using software on a programmable platform instead of implementing on the hardware, so that modulation, coding scheme,

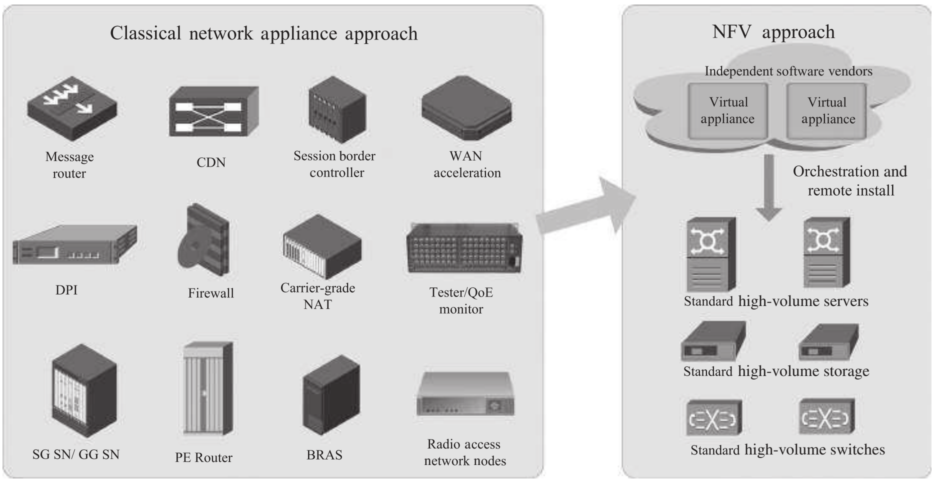
Classical network appliance approach
Message router
CDN
Session border controller
WAN acceleration
DPI
Firewall
Carrier-grade NAT
Tester/QoE monitor
SG SN/ GG SN
PE Router
BRAS
Radio access network nodes
NFV approach
Independent software vendors
Virtual appliance
Virtual appliance
Orchestration and remote install
Standard high-volume servers
Standard high-volume storage
Standard high-volume switches

Figure 1.4 NFV vision.

resource management, and so on can be adaptive to various scenarios. Another famous example is SDN [5, 14]. SDN decouples the control plane and data plane, where the central controller is responsible for path selection based on the global network state. This configuration has many benefits, including cost reduction on routers, fast deployment of network services, capable of supporting UE from various applications, and so on. Nowadays, network function virtualization (NFV) technology enables operators to manage the infrastructure more efficiently [15, 16]. The goal of NFV is to transform the infrastructure from fragmented non-commodity hardware into platform building on common internet technology (IT) servers and storages, as shown in Fig. 1.4 [15]. To maximize the benefit from IT, end-to-end reconfigurable designs should be considered. Besides the NFV utilized in core network, virtualization of RAN, known as cloud radio access network (C-RAN), has been proposed [17]. A C-RAN network brings the baseband units (BBUs) processing resource to a centralized pool so that the resource can be managed and allocated on demand. The centralization of the baseband processing provides more energy-efficient cooling. By virtualizing the baseband processing, new features can be added to the network within months or even a few days, as opposed to years in the traditional infrastructure.

The soft concept should be extended to the air interface as well. The traditional air interface design focused on a “one-size-fits-all” approach to achieve a global optimization or trade-off. However, when it comes to 5G, soft physical layer air interface design is needed to assist with SDN and NFV technologies for providing users with diversified services and consistent quality of experience. To address this issue, the concept of software-defined air interface (SDAI) has been proposed to provide a configurable mechanism to customize air interface design to support different services under different conditions [18]. SDAI involves an intelligent controller to make air interface service oriented, and the multiple fundamental building blocks, such as multiple access, waveforms, modulation and coding, and spatial processing schemes, are case-specific configurable but with unified architecture. The unified architecture and the maximum sharing of foundational functionalities should be utilized as much as possible to ensure high energy and computational efficiency.

1.4 Green vs. Soft: An Unsolvable Contradiction?

For several decades, it has been widely accepted that dedicated hardware like an application-specific integrated circuit (ASIC) is required in communication systems to achieve highly efficient BS operation. The BSs were designed following this approach. If these functions are to be realized using field-programmable gate array (FPGA), which is capable of flexible configuration of baseband algorithms according to different scenarios and radio access technology protocols (e.g., LTE and WiFi), the power consumption would be much higher and green communication is thus difficult to achieve. Therefore, the contradiction between green communication and soft implementation has puzzled both IT and communication technology (CT) engineers worldwide, such that green plus soft design for wireless communication system has been regarded impossible for a long time and has not been investigated seriously.

Generally the contradiction between green and soft is true for a single BS implementation. The softer the implementation of algorithms, the more energy consumption it incurs. If we take the whole communication network into consideration, is the contradiction still valid? Is it possible to achieve green plus soft design if we pool the baseband computation in a central processing unit and conduct soft implementation using FPGA or general-purpose CPU with smart workload allocation based on the temporal and spatial mobile traffic variations? Later on in this book, we will show how green and soft network design can be achieved simultaneously.

1.5 Rethinking Green and Soft 5G Network Design

Characterized by a mixed set of KPIs, such as data rates, latency, mobility, energy efficiency, and traffic density, 5G services demand a fundamental revolution on the end-to-end network architecture and key technologies design. Toward a “Green and Soft” 5G, eight innovative 5G research and development themes have been proposed by China Mobile, including:

1. Rethinking Shannon to start a green journey on wireless systems;
2. Rethinking Ring and Young for no more “cells”;
3. Rethinking signaling and control to make network applications- and load-aware;
4. Rethinking antennas to make BSs invisible via SmarTiles;
5. Rethinking spectrum and air interface to enable wireless signals to “dress for the occasion”;

Cambridge University Press
978-1-108-41732-7 — Green and Software-defined Wireless Networks
Chih-Lin I , Guanding Yu , Shuangfeng Han , Geoffrey Ye Li
Excerpt
More Information

6. Rethinking fronthaul to enable soft RAN via next-generation fronthaul interface (NGFI);
7. Rethinking the protocol stack for flexible configurations of diversified access points and optimal baseband function split between the BBU pool and the remote radio systems;
8. Rethinking big data (BD) analytics in wireless communication systems to facilitate globally optimized resource allocation and scheduling via big-data-enabled network architecture and signaling procedure.

1.5.1 Rethink Shannon

After decades of high-speed development, the scale of ICT, particularly communication networks, is huge enough such that its power consumption is no longer a negligible factor in global energy consumption. Considering a 1,000-times capacity increase by 2020, the power consumption of future networks is not affordable if the network is designed with the current energy-scaling rule.

Classical Shannon theory, a “bible” in the technical communications domain, has been leading the development of communication systems for more than half a century. The extension of Shannon theory from scalar to vector in the early 1990s triggered the invention of the MIMO technology, which brought another golden twenty years of wireless communication systems. The SE and EE relationship is recently explored by rethinking Shannon theory, with a simple mathematical manipulation, for guidance on development of future green communication systems in the next decade. By only considering transmit power over the air that traditional Shannon theory dealt with, a monotonic trade-off between SE and EE always exists, which means that increasing SE will induce an EE reduction. That would not have been very interesting, nor useful. However, in any realistic network operations, the circuit power consumed by the equipment also takes an important part besides the transmit power. This power accounts for a greater and greater share of the total power as the cell becomes smaller and smaller. If taking into account the circuit power, the relationship between SE and EE is no longer monotonic. There is actually a win-win region for EE and SE, which presents a broad R&D field for joint SE and EE optimization [19–21]. It applies in future networks from each individual component technology to network-wide performance evaluation, ranging from the equipment level to the network level.

Diverse traffic fluctuation in the temporal and spatial domains provides another opportunity to rethink Shannon theory, and different scales of traffic characteristics can be well exploited to improve both SE and EE. Network architecture and deployment can be smartly optimized by taking advantage of spatial correlation properties. Resources can be more efficiently managed and allocated by using the small-scale variations of traffic volume. Transmission technology can be adaptively selected or combined in different scenarios to implement EE–SE codesign. In 5G standardization, many new technologies are being studied, e.g., hybrid beamforming for higher frequency bands, non-orthogonal multiplex access schemes, and new waveforms. The EE–SE codesign needs to be taken into consideration.

1.5.2 Rethink Ring and Young

The concept of cellular systems was proposed in 1947 by two researchers from Bell Labs, Douglas H. Ring, and W. Rae Young. Since the first generation of cellular standards, this cell-centric design has been maintained through every new generation of standards including 4G. Toward the timeline of 2020, with the introduction of HetNets and ultra-dense networks (UDNs) [22], multiple layers of radio networks have come into being. Energy consumption, interference, and mobility issues are becoming more serious due to smaller inter-site distance. Diverse types of BSs with different coverage, transmit power, frequency bands, among others are introduced. Traffic fluctuation is more significant than before, taking into account emerging millions of mobile data applications. Therefore, in practical deployment, it is clear that the traditional homogeneous cell-centric design of mobile networks does not match the anticipated traffic variations and diverse radio environments.

The design of user-centric 5G radio networks should start with the principle of "no more cells" (NMC), departing from cell-based coverage, resource management, and signal processing. It should be predicated on the spatial and temporal variation of user demand, rather than a fixed cell-bounded configuration.

Given a great deal of overlapped coverage in a UDN, to alleviate interference, more radio channel information between radio access points nearby should be shared in real-time, and more joint collaboration between neighboring cells is required. Dynamically for each user, the available radio resources from multiple access points could be jointly scheduled, and the selection of control plane (CP)/user plane (UP) and UL/DL channels, respectively could be done separately.

Centralized mobility control across different cells is also essential to reduce handover interruption delay. Besides that, multi-connectivity is viewed as a promising way to realize high throughput, ultra reliability, and seamless mobility. Multi-connectivity control and user plane anchor require centralized processing across multiple cells.

In addition, enabled by SDN and NFV, multi-RAT convergence and centralized BD processing are also motivating centralized processing across multiple cells.

A macro BS, utilizing LTE evolution or a new RAT at lower frequency, provides wide coverage and serves as a signaling BS while small cells at higher frequency, such as millimeter wave (mmWave), aim at boosting throughput and offloading traffic. Furthermore, to reduce the CAPEX/OPEX of small cells, by considering smaller coverage, supporting fewer users with low mobility, more relaxed synchronization requirement, and smaller time and frequency selective fading, "data-only carrier," with on-demand system information without broadcasting overhead, can be implemented to reduce interference and energy consumption. Macro cells can also help small ones in discovery, synchronization, measurement, etc.

With the emergence of C-RAN, many technologies toward realization of the concept of NMC can be facilitated. By taking into account differentiated fronthaul conditions, RAN can be split into a central unit (CU) and a distributed unit (DU), where the CU is in charge of centralized collaboration and user plane anchor, and the DU is responsible for radio scheduling and transmission.

1.5.3 Rethink Signaling and Control

As the proliferation of mobile internet continues, new services and applications appear at a fast pace. Some have exhibited orders of magnitude higher overhead over-the-air than more traditional services, since signaling/control of current networks is “connection-oriented.”

In the 5G era, the user and traffic characteristics will be much more diversified, and the resource-contending environment will be more complex [23]. Therefore, more intelligent and adaptive signaling/control mechanisms are desired for 5G networks to achieve low-cost transmission with high signaling efficiency. Thus, 5G over-the-air signaling/control should be an intelligent combination of both connection-oriented and connectionless mechanisms. It should be also application aware, load condition aware, and user status (e.g., mobility) aware.

A new lightweight state, besides IDLE and CONNECTED, should be introduced to support “connectionless” mode. Under such state, the end-to-end (E2E) connection shall be resumed quickly without starting from scratch, so that access delay can be reduced significantly, and signaling overhead can also be reduced accordingly.

In addition, RAN signaling and control should not be limited to RAN protocol layers. Cross-layer optimization between RAN and high-layer applications seems to be a promising technology trend. RAN could be enhanced to “smart RAN” with service awareness without impairment of user privacy to improve users’ quality-of-experience (QoE), for example, application level adjustment with radio channel fluctuation, and differentiated RAN L2/L3 treatment with service awareness.

Furthermore, the signaling and control should be slice aware and tailored for service requirements of diversified slices. The mobile networks shall be able to provide differentiated slices with customized signaling/control, where the differential access control, network entities, mobility management, security control, etc. are totally on demand. For example, during the low load period, a slim air interface can be configured to achieve low cost [24]. Customized signaling/control for differentiated network slices shall be designed for different contexts (user, service, network circumstance). More importantly, a network framework is required for signaling/control allocation and network function orchestration. SDN is extremely suitable for such a signaling/control framework. It provides a flexible and centralized control framework, and its open programmable interfaces also make it scalable to support new services. Moreover, with the centralized SDN framework, more contexts should be collected, and the big-data-enabled processing will be performed better.

1.5.4 Rethink Antenna

Targeting significant capacity enhancement in 2020, the 5G network is expected to be ultradense with massive antennas deployed either in a distributed or centralized manner. Theoretically, massive MIMO or large-scale antenna is expected to significantly improve network capacity and reduce the inter-cell and intra-cell interference, hence they may enhance both the SE and EE. However, to accommodate a few hundred antennas and transceiver chains all on one infrastructure in a traditional cell site is very