3 Green and Soft Network Design

3.1 Green and Soft Wireless Communication Network Design

The 5G network is anticipated to be reconfigurable with a software-defined network. A soft network is envisioned to bring agility into the implementation of each network element from core network to access network, as well as the building blocks of the air interface. The network function and resource virtualization should be the core of a soft network. It decouples software and hardware, control and data, uplink and downlink to facilitate a converged network, as well as information technology and communication technology convergence, multiple radio access technology (RAT) convergence, radio access network (RAN) and core network (CN) convergence, content convergence, and spectrum convergence. This enables a super-flat architecture that achieves cost-efficient network deployments, operation, and management.

In a soft network, the computing, storage, and radio resources are virtualized and centralized to achieve dynamic and user-centric resource management, matching service features. Soft networks are expected to build on a telecom-level cloud platform to enable network-as-a-service with the features of open network capability and network sharing. This makes it possible to achieve network flexibility and scalability and provides users with a massive variety of services and consistent quality of experience. Meanwhile, this will lead to much greener network designs and operations from the bottom up.

In this chapter, the design principle of the end-to-end (E2E) network architecture for 5G is first presented, including the considerations on the core network, transport network, and RAN. Then the cloud radio access network (C-RAN) architecture is elaborated as an enabling technology for a green and soft 5G network. With the development and the increasing convergence of big data (BD) analytics, communication technologies, and information technologies, BD-enabled wireless communication networks are motivated, which are anticipated to provide satisfactory services in diversified scenarios with globally optimized resource allocation and maximal extent of software configurations of the E2E network architecture. BD-enabled wireless network design is further investigated, including the potential impact on the network architecture, protocol stack, signaling, and physical (PHY) layer operation. Finally, the benefits of applying BD analytics to the wireless communication network are examined, for the case of mobility management and cross-layer transmission control protocol (TCP) optimization.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

3.1.1 E2E Network Architecture for 5G

Besides the performance improvement of mobile network and support of highly diversified applications [1–3, 22, 23], the most important requirements of E2E 5G network architecture design are green and soft. Toward green network functions, interfaces, and protocols can be further simplified and converged, e.g., LTE and 5G network can converge at the RAN side to avoid excessive handover overhead; low-cost deployment, especially under ultra-dense network, efficient flow forwarding, and optimized flow routing, can be implemented to reduce E2E latency and avoid congestion. Besides, part of baseband processing can be centralized based on different fronthaul conditions to improve collaboration capability and pooling gain. In the meantime, the latency can be significantly reduced especially for URLLC services [4–8] by introducing mobile edge computing. Toward soft networks, the transition is inevitable from fixed network entities and deployment, as well as static connection in 2G/3G/4G to NFV (network function virtualization) by introducing functions virtualization and flexible E2E function orchestration to implement dynamic configuration, flexible connections, etc. Besides, operator's revenue can be improved via providing customized services per slice granularity.

5G E2E network architecture has the following characteristics, compared with 4G:

- E2E network slicing: to satisfy diverse requirements from vertical industries, E2E network slices are required to provide guaranteed quality and customized services from UE, RAN, CN, transport network, etc.
- Service-based and componentized core network: service-based function design can facilitate flexible customization and aggregation of network functions. In addition, user plane can be simplified to achieve efficient forwarding by splitting control and data forwarding.
- Flexible and smart RAN: centralized management and collaboration gain from RAN-side can be achieved by separating the central unit and distributed unit. With non-stand-alone deployment, 5G new radio and LTE evolution can be converged at RAN-side [10] to maximize legacy network investment. In addition, by introducing smart service processing on the RAN side to enable mobile edge computing, E2E latency can be greatly reduced, and more local optimization is possible.

In summary, 5G overall architecture design principles of RAN, CN, and transport network perspectives are as follows: for RAN, a two-level architecture with CU/DU split to support flexible deployment is preferred; for CN, a service-based network architecture is motivated; for transport, the fronthaul needs to support CU/DU split, low latency, and high throughput. In addition, the following key issues may be related to E2E architecture and will be further described in the later sections: E2E network slicing, flow-based QoS, non-stand-alone deployment, and edge computing.

3.1.2 Next-Generation Core Network

The traditional core network [14] is generally based on the tightly coupled design of network entities functions, interfaces, and signaling flows between entities. However,

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

this concept does not adapt well with rapid network upgrading, fast time-to-market feature introduction, and system performance scalability. The telecom industry learns from the internet industry the service-oriented architecture (SOA) [31], a similar concept, which is relatively loosely coupled and popular in internet architecture design and can be introduced in the 5G core network design.

A 5G core network is service-based to achieve a flexible combination of network functions, rapid time-to-market, agile usage, and independent scalability. The service-based architecture has the following two major characteristics:

- Micro-service-based network function design: 4G network designs are based on dedicated hardware and network functions are tightly coupled; however, the 5G network is redefined to decouple the network functions of the control plane, which allows for flexible combinations and independent evolution.
- Service-based interface model: Lightweight communication protocol is enabled between network functions to allow easy function invocation; network functions may be invoked on-demand via standard lightweight protocols, which tend to improve efficiency and reduce development complexity.

An example of 5G system reference architecture excerpted from 3GPP TS23.501 [13] is illustrated in Fig. 3.1. A number of network functions, such as NSSF (network slice selection function) and NEF (network exposure function), are defined as types of services, which can be easily tailored for diversified service requirements (e.g., invoking certain functions or expanding capacity on-demand).

These descriptions are all mainly focused on the RAN. The concept of soft design should also be reflected in the evolution of core networks (CNs).

In current LTE networks, there are still many challenges and obstacles for operators to maintain and/or upgrade their services. For example, the EPC entities like MME, SGW, PGW, and PCRF are typically based on customized hardware. This is always cost-inefficient and inflexible for network management. Recently, software-defined networks (SDNs) [27–30] and NFV are being generally identified as the most promising technologies for next-generation CN architecture. SDN decouples the control plane and

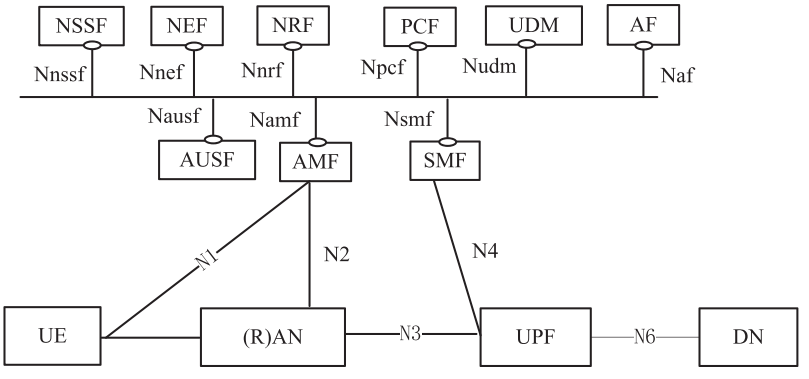

Figure 3.1 5G system reference architecture.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

the data plane to reduce the complexity of distributed networking control protocols by using a centralized controller. NFV decouples the hardware and the software, where proprietary hardware network elements are replaced with virtual applications running on low-cost, general-purpose hardware platforms. The integration of SDN and NFV facilitates the deployment of cloud computing with network virtualization, allowing for ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources. This brings flexibility, reconfigurability, service elasticity, and vendor independence, and also shortens the time-to-market of new applications and services.

Meanwhile, the emergence of SDN and NFV is a great driving force of the development of network slicing technique, which has been envisioned as a key element to meet the diverse demands of 5G use cases and underlying cost requirements. In network slicing, CN is considered the most critical part of providing network services for tenants and their end users. To meet the diversified demands of vertical industries, service customization and on-demand deployment are the key concepts that need to be reflected in CN design. The softness and flexibility inherited in SDN and NFV will enable operators to set up services quickly, and move them around as virtual machines in response to network demands.

3.1.3 Next-Generation RAN

Under ultradense networks, collaboration and centralized mobility control need to be reinforced, since interference and handover signaling overhead may be rather severe. Multi-connectivity seems to be an important way to realize high throughput and ultra reliability. Besides, a number of factors, such as multi-RAT convergence, centralized SON, and NFV/SDN may motivate a new design of RAN architecture, i.e., CU (central unit) and DU (distributed unit) split with an open interface to support interoperability. More specifically, as shown in Fig. 3.2, the functionalities of CU and DU are as follows:

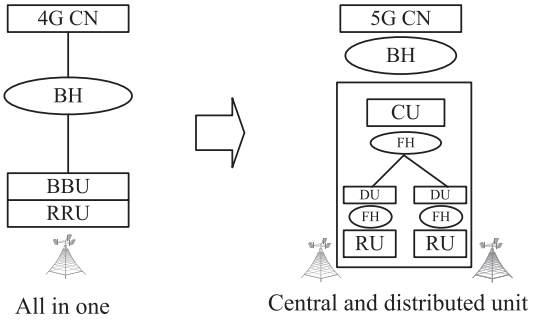

Figure 3.2 From all in one to central and distributed.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

- CU mainly deals with non-real-time processing of RAN protocol stack functionalities, which is composed of L3 (e.g., RRC) and partial L2 [16]. CU is required to be centralized and enhanced to support multi-leg management, such as multi-leg reordering, flow control, retransmission, etc.
- DU consists of L1 and partial L2, which is latency-sensitive and mainly performs upon a TTI level, such as ARQ in RLC, and segmentation upon MAC request.

A number of split options have been discussed in 3GPP RAN3 [12]. Option 2 (PDCP and RLC) has been selected for standardization of high layer split [15] and low layer split is still under discussion.

CU is the RAN controller and is in charge of data distribution, which is able to implement multi-connectivity, seamless mobility management, and efficient spectrum usage, which is not a counterpart like RNC in 3G network.

Regarding supporting dual/multi-connectivity, CU is able to serve as an anchor to support data distribution between multiple DUs, which is beneficial to reduce backhaul pressure for data distribution between base stations without anchor points. Additionally, it is also easier to plan the bandwidth of transport network. Looking forward to 5G, CU can be the common anchor for 4G and 5G, which can accelerate data distribution between multi-RATs. In addition, if the transport network between CU and DU is reliable and latency is under millisecond level, multicell radio resource management (RRM) can be introduced in CU to implement relatively rapid radio resource scheduling and coordination of multicells and multi-RATs.

On one hand, regional rapid RRM in CU may play an important role in future networks since it is expected that ultradense networks would be deployed with condensed cell radius and highly overlapped coverage from multiple cells. Under such cases, RRM by the granularity of cells and TTI may not be optimal anymore.

On the other hand, regarding the support of E2E network slicing, it is crucial to implement air interface slicing to satisfy SLA (service level agreement) requirements of diversified services since air interface may be the bottleneck of E2E QoS guarantee. CU and DU split can by nature support E2E slicing. CU can be enhanced to support slice isolation. DU is able to support differentiated configurations, flexible air interface scheduling with customized slicing strategies.

The deployment of CU and DU is required to consider a number of factors, such as transport network conditions, RAN equipment complexity, pooling gain, collaboration gain, and maintenance costs. If the transport network between CU and DU is ideal with high throughput and low latency, such as dark fiber, real-time processing of RAN protocol stack functionalities can be centralized to achieve maximum collaboration gain. However, if the transport network is restricted with limited bandwidth and a long latency, only non-real-time processing is centralized. In addition, the deployment can be adjusted based on service requirements. For example, CU functionalities can be further split into CU-C (central unit – control) and CU-U (central unit – user plane). CU-C can be placed to support large-scale radio resource management and control, and CU-U can be deployed close to UE to support low latency requirement.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

3.1.4 Next-Generation Transport Network

5G transport network design is aimed at providing large bandwidth and low latency [12]. Firstly, the fronthaul transport network needs to satisfy the transport requirements from RAN CU and DU split architecture. If high layer split is adopted, delay-tolerant and relatively low-throughput fronthaul transport network is needed. If low layer split is adopted, delay-sensitive and high-throughput fronthaul transport network is needed. In addition, synchronization with high precision is also required to support CU and DU separation. Secondly, the advantage of SDN can be taken to realize flexible configuration and even slicing of network resources.

Meanwhile, due to potentially explosive traffic demands, locally processed data in particular may increase exponentially with the introduction of mobile edge computing, layer 3 IP switch capability would be pushed down to transport aggregation ring to enable local IP flows processing.

3.1.5 Key Issues of E2E Network Architecture

Non-Stand-Alone (NSA) and Stand-Alone (SA) Deployment

In order to take advantage of good coverage capability from well-established LTE low-frequency networks, and in the meanwhile 5G new radio capabilities, non-stand-alone (NSA) [17] deployment is introduced in the first phase. Under this mode, UE relies on LTE BS to provide control plane signaling between UE and CN [25]. Currently, there are two options, as shown in Fig. 3.3:

For option 1, LTE provides signaling connection to 5G NR base station via 4G CN (EPC, evolved packet core). 5G NR (new radio) base station is expected to only provide user plane data transmission to boost capacity. The capability to support dual connectivity of user planes [26] via LTE and 5G NR is needed for terminals.

Option 2 is different from option 1 in that LTE is expected to evolve to support NG Core (next-generation core), which supports new functionalities, such as service-based architecture, network slicing, flow based-QoS, etc. The signaling to support NR data transmission is still anchored to LTE, and the user plane of NR can be either connected to LTE or NG core for boosting capacity.

Regarding the upgrade cost, for option 1, upgrade of 4G RAN and EPC is required to support dual connectivity [21]; for option 2, upgrade of 4G RAN, to support dual connectivity, and connection to NG core is required. EPC upgrade is not required. For both options, if the data plane is split at RAN, upgrade of 4G BBU hardware is required to support 5G NR high throughput. The cost may be relatively high, but higher performance of data aggregation with RAN control may be achieved. In contrast, if data split is expected at the core, 4G BBU hardware upgrade is not required.

NSA mode is expected to be an interim phase and eventually would be replaced by SA deployment. For SA mode, the NR base station works independently and UE connects with the NG Core (e.g., registration and authentication) via NR base station. There are still two options for SA mode: options 3 and 4 in Fig. 3.4. For option 3, a 5G NR system

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

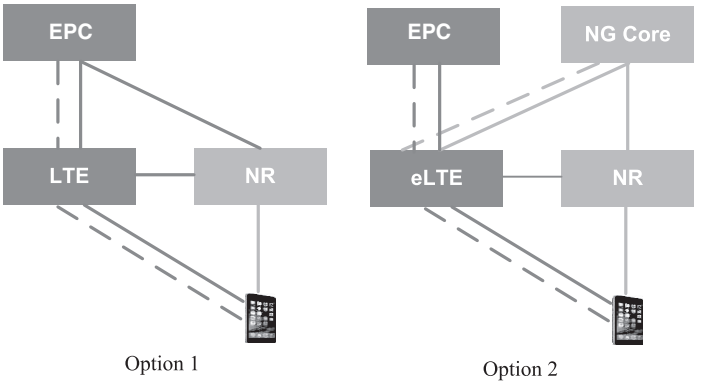

Figure 3.3 NSA options.

Figure 3.4 SA options.

is required to interwork with LTE from irrespective core networks, i.e., NG core and EPC. For option 4, a 5G NR system is expected to interwork with LTE via NG Core.

For option 3, 4G RAN upgrade is not required, and only EPC is required to be upgraded to support interworking with a 5G NR system. The workload for a network upgrade is relatively low; however, interworking performance may not be as good as expected.

For option 4, both hardware and software upgrades for 4G RAN are needed, but EPC upgrade is not required to support the connectivity to NG core.

E2E Network Slicing

It is foreseeable in 5G that E2E network slicing is indispensable in satisfying emerging diversified services. A network slice refers to a group of logical network functions to satisfy specific quality requirements from vertical industries, which require physical and virtualized resources as needed, including RAN, CN, and even the support of the transport network, or IP network.

From the NGMN paper "Description of network slicing concept" [11], the network slicing concept consists of three layers: service instance layer, network slice instance layer, and resource layer. The service instance layer represents the services (end-user services or business services) that are to be supported. Each service is represented by a

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

service instance. Typically services can be provided by the network operator or by third parties. In line with this, a service instance can either represent an operator service or a third-party provided service.

A network operator uses a network slice blueprint to create a network slice instance. A network slice instance provides the network characteristics that are required by a service instance. A network slice instance may also be shared across multiple service instances provided by the network operator.

The network slice instance may be composed of none, one, or more subnetwork instances, which may be shared by another network slice instance. Similarly, the subnetwork blueprint is used to create a subnetwork instance to form a set of network functions, which run on physical/logical resources.

To implement "network slicing as a service," it is important to design optimum network slicing solutions for each specific use case from vertical industries, including network topology, functions, and protocols. Logically independent network slices may coexist in a shared physical infrastructure. Some functions may be shared by multiple slices, while others need to be customized via parameter reconfiguration or redesign. As shown in Fig. 3.5, an example of RAN slicing is described. Some common network functions (e.g., RRC) are implemented, and slice-specific network functions are available as well, for example, RRC, PDCP, RLC specially designed for URLLC and mMTC. For the low layer, the radio resource is shared by all types of slices to achieve maximum spectrum efficiency.

More specifically, some key issues should be addressed to implement an efficient RAN slice [9].

Efficient radio resource management between slices is crucial for improving spectrum efficiency. Since network physical resources (e.g., computation and storage) could be virtualized and customized for dedicated services, radio resources would similarly be virtualized as time/frequency/space granularity from the following perspectives.

- Spectrum planning: relatively static spectrum allocation (e.g., as a granularity of carrier) for different slices could be implemented. In such a way, sliced QoS

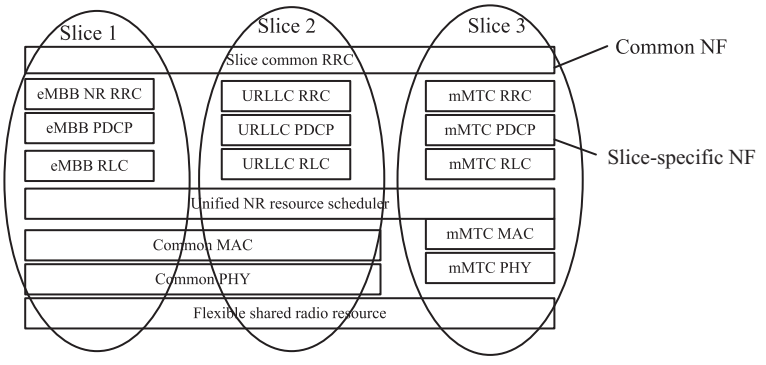

Figure 3.5 An example of RAN slicing.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

can be easily guaranteed as expected since radio resource blocks are reserved beforehand, however, the degradation of spectrum efficiency is unavoidable. In addition, the spectrum adjustment is allowed and could take hours or days.

- RT (real-time) scheduling: To maximize spectrum efficiency, radio resource management could perform as a granularity of TTI level, just like common MAC scheduling for different slices. As a result, the radio resource collision may probably happen in the case of bursty traffic in some slices to degrade SLA of other coexisting slices.
- Non-RT scheduling: Similarly, the radio resource management could perform on a larger time scale, for example, a few hundreds of TTI level. The collision probably could be reduced accordingly.
- Access control: Radio resource allocation could also be done via access control per UE level. Some UEs in one slice may be barred from accessing the network since more radio resources are required for another slice.

Different ways of radio resource management may be adopted for diverse scenarios by taking into account network deployment and service requirements. In addition, different levels of adjustment shall be allowed for these ways. BD analytics could be used to make an optimized estimation of radio resource fluctuation so that the corresponding adjustments could be performed in a timely way.

Since the KPIs of eMBB, URLLC, and eMTC are highly diversified and most likely may not be supported over all cells of the RAN. Thus it is important to consider slice availability all cross different cells especially for initial access, handover, and secondary node selection in dual/multi-connectivity.

Flow-Based QoS

4G QoS granularity is based on the E2E bearer, which is too rough to satisfy 5G flow-based service requirements. Moreover, the overhead of E2E bearer establishment or modification is too high to meet rapidly varying QoS requirements of 5G. Flow-based QoS is introduced in 5G to implement finer granularity and faster control.

QoS flow is the granularity of QoS differentiated treatment. During session setup, the CN is able to notify RAN and UE of flow-based QoS profiles, so that flow level treatment can be implemented in RAN and UE.

An in-band QoS control mechanism is also introduced to allow real-time QoS adjustment [19]. In-band QoS marking in a user plane can be modified without signaling involvement, so that flow-based QoS treatment can be reinforced at any time to facilitate real-time processing, which avoids excessive signaling interaction.

Mobile Edge Computing

With the rapid growth of data traffic and diversified services booming, the application of mobile edge computing (MEC) [26] is becoming more and more popular [20]. For example, in the short term, high data traffic demand in enterprises and campuses is expected. In the midterm, low latency and high throughput services, such as AR

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

(augmented reality) may explode in the mobile network. In the long term, auto-drive may become reality with the support MEC.

In general, MEC is adopted not only for performance improvement, such as latency reduction, but also edge intelligence enhancement to boost new business models.

For enterprises, MEC can be used to replace WLAN networks to provide reliable, high-speed, secure, low-cost data access for mobile office automation, industrial control, and IoT (internet of things). For campuses, in order to avoid excessive occupation of a backhaul transport network, MEC can allow users to directly access local networks without traversing the operator's core network.

In the meanwhile, the platform deployed close to UE can process local data and provide user profiles to facilitate more business opportunities. Video surveillance is another good example of MEC's application. Usually most video content is useless and viewed as a waste of backhaul traffic. If a MEC platform can process the local video content with video and picture recognition capabilities, not only can the backhaul pressure be relieved, but efficiency can be improved as well. AR and auto-drive service can also be viewed as typical scenarios of MEC deployment. Normally, AR and auto-drive are required to implement real-time interaction. MEC is crucial to realize low latency and intelligent local data processing.

Small scale of data centers can be deployed at RAN side to offload specific service flows. In the meantime, some computation-intensive processing can be implemented at the edge in a distributed way to reduce backhaul traffic and relieve the processing burden of centralized data centers. For example, image and video reorganization is being processed at the edge, and only recognized information is required to be backhauled to the central server [26].

Meanwhile, RAN network capabilities can be visible to the third-party applications hosted at the edge to improve its performance or incubate new business models. For instance, the information on UE available bandwidth can be gathered and supplied to the application server to enable adaptive transmission, e.g., TCP congestion window adjustment, or for the video server to adjust coding rate under varying wireless conditions.

3.1.6 Summary

To meet the challenging requirements of 5G services, 5G network architecture is anticipated to have the following distinctive characteristics: service-based core network, flexible RAN deployment with CU/DU split, network slicing with flexible orchestration of redefined network functions, flow-based QoS control, edge computing, etc. In 3GPP standardization, RAN CU/DU high-layer split has already been adopted for specification, and low-layer split options and further split between CU control plane and CU user plane are being studied. Regarding slicing, E2E networks (including CN, RAN, and UE) is required to be aware of E2E slicing for slice SLA guarantee. Service-based CN is agreed as the only option for CN evolution. Edge computing is also being standardized in SA2, and context awareness in RAN to accelerate content delivery has also been

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

studied in RAN3. The concept of SDN/NFV has been well reflected in 5G network architecture design, which allows 5G network to progress towards "Green" and "Soft."

3.2 C-RAN: Revolutionary Evolution of RAN

3.2.1 Introduction

The telecom industry has been witnessing a traffic explosion in recent years. It is estimated that internet traffic is expected to increase over 1,000 times by the year 2020, with over 50% of the traffic volume in file sharing. Unfortunately, operators will not see a proportionate increase in revenue. Instead, mobile operators have to invest in more infrastructure just to keep up with such data explosion, significantly increasing the total cost in addition to increasing the total cost of ownership (TCO), while complicating network maintenance, as there are legacy 2G, 3G, and 4G components all coexisting with each other. In addition, system upgrades will become even more challenging when 5G is introduced [32].

The traditional RAN architecture is facing various challenges in the 4G era and beyond. First, traditional network deployment usually requires a separate room per site with supporting facilities such as air conditioning to accommodate the base station (BS) or baseband unit (BBU). This form of deployment is becoming increasingly difficult since the available real estate is becoming scarcer and rental costs are increasing. Furthermore, it could be foreseen that this issue would become more severe when heterogeneous networks with a high density of small cells begin to prosper.

Second, interference problems in the current LTE networks are much more severe than in 2G and 3G networks due to a larger number of small cells in order to facilitate higher data capacity. In order to mitigate this interference, collaborative radio techniques, such as coordinated multipoint (CoMP) [33], have been proposed. However, efficient CoMP algorithms, such as joint transmission (JT), cannot achieve maximum performance gain with the traditional X2 interface with LTE architecture due to high latency and low bandwidth [34, 35]. Furthermore, in 5G the network would be much denser than 4G and therefore the interference issue is more critical.

Last but not least, power consumption is a great concern for operators as both the carbon footprint and energy costs of the network increase. From [36, 37], a large percentage of power consumption in mobile networks comes from RAN. As a result, saving energy in the network's RAN directly lowers the operation expense (OPEX) of the network.

Centralized, collaborative, cloud, and clean RAN [36, 38] (C-RAN), proposed by China Mobile Communications Corporation (CMCC), is a new type of RAN architecture to help operators address the aforementioned challenges. A C-RAN system centralizes different baseband processing resources to form a single resource pool such that the resource can be managed and dynamically allocated on demand. C-RAN has several advantages over traditional BS architecture, such as increased resource utilization, lower energy consumption, and decreased interference due to better support for CoMP implementation.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

There have been many studies on C-RAN in the literature [39, 40]. In this chapter, we will briefly recall the original definition of C-RAN as well as its features and advantages. Then we will focus on the evolution of C-RAN, i.e., how C-RAN is evolving to become the essential element of 5G to meet its diverse requirements.

3.2.2 C-RAN Basics

C-RAN was proposed as a key RAN architecture by CMCC in 2009. The basic idea of C-RAN is to centralize the processing resources (e.g., baseband processing resources) in the pool and, further, virtualize them to realize on-demand allocation. Its original definition includes the following four key features.

- Centralization: Instead of requiring one equipment room for each base station, as in the traditional deployment, a certain number of BSs would be centralized in a bigger room in C-RAN. With centralization, first, site selection becomes much easier, since there is no need to find many sites. This is particularly important given that LTE systems require many more sites than 2G/3G due to higher frequency, and the expense on real estate is getting higher and higher in current society. Furthermore, network construction time could be reduced and network deployment could be sped up. In addition, using a big central room to accommodate several BSs means that the BSs could share the same facilities, such as power, air conditioning, etc., which then contributes to the saving on OPEX. In fact, the benefits from centralization have been continuously verified in CMCC's commercial networks. For example, it was observed that power consumption could be reduced by 41% in one trial due to shared air conditioning.
- Cooperation: The idea behind it is to use high-speed and low-latency switches to connect the BSs in the same central room so that the BSs could cooperate with each other. In this way, it is expected the system performance could be improved with the support of cooperative technologies, such as CoMP. C-RAN's capability to facilitate CoMP technologies has been verified in commercial networks, where improved system performance has been observed.
- Cloudification: Instead of traditional equipment that is developed based on specialized hardware such as a digital signal processor (DSP), field programmable gate array (FPGA), etc., the ultimate C-RAN system is targeting to adopt general-purpose hardware, e.g., standard IT servers to realize the mobile communication functionalities. In fact, the idea of cloudification is in line with the philosophy of Network Functions NFV.
- Clean system: C-RAN systems are expected to enjoy much more energy saving than the traditional architecture for several key reasons, namely, facilities centralization, which helps save energy with such tactics as shared air conditioning, and the adoption of NFV [41].

C-RAN realization is a stage-by-stage process with different features realized at different stages. Centralization is the first step, while cloudification is the ultimate goal of C-RAN systems.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

3.2.3 Evolution of C-RAN towards 5G

The concept of C-RAN is evolving in the past few years as research and technology progresses. In particular, the concept of centralized unit/distributed unit (CU/DU) function re-split and next-generation fronthaul interface (NGFI) are introduced in C-RAN. Combined with such new concepts and new features, evolving C-RAN has become the essential element of 5G.

5G BBU will be further divided into a CU and a DU. The principle of CU/DU functional re-split lies on real-time processing requirements from different functions. A typical CU mainly includes non-real-time RAN high-layer processing, functions migrating from CN and MEC services. Accordingly, a DU is mainly responsible for physical layer processing and real-time processing of layer 2. In order to lessen transport requirement between the RRU and the DU, partial physical layer processing can be moved from the DU to the RRU. From the equipment point of view, CU equipment can be developed based on general purpose platform, which supports RAN functions, functions migrating from the CN and MEC services. DU equipment can be developed based on a customized or hybrid platform, which supports intensive computing. With NFV infrastructure, system resources, including the CU and the DU, can be orchestrated flexibly via management and orchestration (MANO), a SDN controller, and traditional operating and maintenance center (OMC), which could support operators' requirements on fast service rollout.

In order to solve the transport challenges among the CU, the DU, and the RRU, NGFI [42, 43] is proposed. As shown in Fig. 3.6, the NGFI switch network provides the connection between the CU and the DU, as well as between the DU and the RRU. With the help of the NGFI, the CU and the DU can be flexibly deployed according to multiple

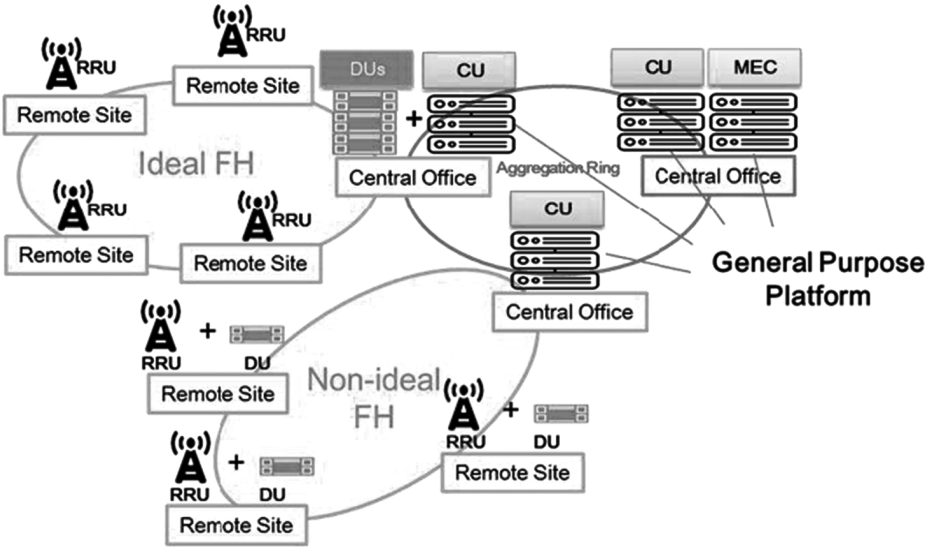

Figure 3.6 CU/DU and NGFI-based C-RAN architecture.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

scenarios. In case of ideal fronthaul, the deployment of the DU can also be centralized, which could support physical layer collaboration. In case of nonideal fronthaul, the DU could be deployed in a distributed way. Therefore, the C-RAN architecture based on the NGFI supports not only DU centralized deployment but also DU distributed deployment.

The concepts of the CU-DU structure and the NGFI will be detailed in the following sections of this chapter.

5G C-RAN is based on the CU/DU architecture, the NGFI, and NFV infrastructure, which has been viewed as a promising 5G RAN architecture. Compared with 4G C-RAN, the main features of 5G C-RAN are still centralization, collaboration, cloudification, and clean (green), with each evolving in the context of 5G.

- Centralization: For 4G C-RAN, all the baseband functions are centralized in the central office. With the introduction of CU/DU and NGFI concept, centralization in a 5G C-RAN has two-fold meanings. First, BBU equipment could be physically centralized, which has been demonstrated to have such advantages as fast network deployment, facility sharing, etc. Second, from the function perspectives, with CU/DU re-split, partial high layer functions could be realized on the same central platform.
- Collaboration: There are two kinds of collaboration in C-RAN. First, the central BBU site serves as a wireless service anchor of the control plane and the user plane, which could support multicell high-layer collaboration, multi-connection, seamless mobility management and cooperative spectrum sensing. Second, when both the CU and the DU are centralized in the center BBU site, the SE (spectral efficiency) of cell edge and average throughput can be improved further by introducing physical layer collaboration, such as CoMP, D-MIMO, etc.
- Cloudification: There are two key aspects of the concept of cloudification. One is function abstraction. The other is decoupling of processing resources and applications. Traditionally, processing resources are allocated to a single BS. For C-RAN cloud, processing resources would be allocated in a resource pool, which is good for processing resource multiplexing, reducing system cost, and flexible network function deployment. For example, MEC, which is viewed as one of cloud RAN's flexible deployment use cases, will be well supported in C-RAN. Moreover, the radio resource can also be abstracted as a resource in the resource pool. Decoupling of the radio resource and radio access technology (RAT), on-demand network capacity adjustment, and service customization can be supported by C-RAN. For example, a specific radio resource can be configured for a designated area in order to meet the requirement of group customers. In general, processing resources and radio resources can be dynamically adjusted according to traffic load, user profile, and service requirements in C-RAN. Therefore, operators' requirement for fast service deployment will be better supported via C-RAN on-demand network capacity.
- Clean (Green): Based on centralization, collaboration, and cloudification, the number of BBU sites, air conditioning, and other facilities can be reduced, and

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

the TCO can be saved by centralization. Moreover, on-demand network capacity and processing resources adjustment are supported. Therefore, overall network efficiency is improved, and the clean (green) target is reached.

3.2.4 NGFI: Next-Generation Fronthaul Interface

Traditional 2G/3G/4G BSs consists of BBUs and RRUs with fibers to connect each other. The link between a BBU and a RRU is called fronthaul (FH). With centralization, different BBUs are clustered in the same room. Thus, from the central room, a lot of fibers would go out to connect the RRUs in the remote sites. The more BBUs are centralized, i.e., the larger the centralization scale is, the more fibers are required. It is unfortunate that fibers have become scarce. The FH issue has become a key obstacle to C-RAN centralization for operators who are lacking fiber resources. There has been some study on the FH solution in literature, for example [36], with several proposed schemes including various compression techniques, wavelength division multiplexing (WDM), optical transport networks (OTN), microwave transmission, and so on. In the later part of this book, we will present field trial results to verify the feasibility of passive WDM FH solutions. In general, it is widely agreed that WDM-based FH solutions are mature enough to effectively save fiber consumption in support of C-RAN large-scale deployment for 2G, 3G, and 4G systems.

When it comes to 5G, the FH issue becomes much more severe and more challenging, given some key 5G features and technologies. One typical example is the massive MIMO technology. In 5G massive MIMO with at least 64 antennas and 100 MHz bandwidth is expected. In this case, the required FH data rate between the BBU and the RRU would be around 100 Gbps, which is unaffordable for even dark-fiber FH connections due to the high cost of 100Gbps optics modules. If C-RAN is further taken into account, the problem would be more critical. There must be a revolutionary design on the FH solution to replace traditional CPRI interface. In fact, this is a consensus in the industry and the answer lies on a new concept, called next generation fronthaul interface.

The basic design principle for NGFI is to repartition the baseband functions within the BBU so that partial baseband functions would be moved from the BBU to the RRU side. As the result, the BBU has been divided into two logical parts and the NGFI connects them.

It is clear the NGFI is depending on used function split schemes, yet from a design perspective, the ideal NGFI is expected to have the following features [43].

- Its data rate should be traffic-dependent and therefore support statistical multiplexing.
- The mapping between the BBU and the RRH should be one-to-many and flexible.
- It should be independent of the number of antennas.
- It should be packet-based, i.e., the fronthaul (FH) data could be packetized and transported via packet-switched networks.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

The key way towards NGFI is the function split between the BBU and the RRU. There has been an extensive study on the comparison for various split options. Please refer to Fig. 3.8 in the next subection for an overview. It is not until recently that 3GPP has finally decided to adopt split option 2 as the split scheme. Option 2 is the split between the packet data convergence protocol (PDCP) and the radio link control (RLC) layer. In other words, with option 2, the PDCP functions would remain inside the BBU, while the RLC layers, together with those layers below the RLC, such as MAC and PHY, would be moved to the remote side and become a new entity. There are more details regarding the split in the next part (Section 3.2.5) of the CU-DU structure.

Function splitting is just the first step for the NGFI. When it comes to the FH networks in the context of C-RAN, there is a radical change compared with the original WDM or other existing FH solutions. Thanks to the packet-based features, it is expected to use packet switching networks to transport the NGFI packets when the ethernet can come into play. Thanks to its ubiquity, low cost, high flexibility, and scalability, the ethernet should be adopted as the NGFI FH solution. There are several benefits. First, an ethernet interface is the most common interface on standard IT servers, and the use of ethernet makes C-RAN virtualization easier and cheaper. Second, the ethernet can fully make use of the dynamic nature of the NGFI to realize statistical multiplexing. Third, flexible routing capabilities could also be used to realize multiple paths between the BBU pools and the RRH.

The main challenges for the ethernet as an FH solution lie on the latency. Traditional ethernet would have the latency capacity of several dozen to several hundred milliseconds, which is unacceptable for 5G scenarios. Low latency is one of the key features of 5G. Not to mention the ultrareliable low-latency communication (uRLLC) application, which requires 1 ms E2E latency. Even for eMBB, the user-plane latency requirement is around 10 ms, which, using traditional ethernet, is hard to meet.

Other challenges include high timing and synchronization requirements imposed by the NGFI interface. Although the exact NGFI has so far not been specified, it is possible that the NGFI may keep some requirements of the CPRI, such as synchronization requirements. The allowable radio frequency error for a CPRI link is 2 parts per billion (ppb). Synchronization is another concern. In order to meet the timing requirements, both the BBUs and the RRUs should be perfectly synchronized, which therefore requires a very accurate clock distribution mechanism. Potential solutions may include any combination of the global positioning system (GPS), IEEE 1588, and synchronous ethernet (Sync-E).

3.2.5 CU-DU Architecture for 5G

CU/DU Definition

The development and standardization of 5G have brought new challenges and opportunities to RAN. Since 5G envisions supporting much wider service types than 4G, the current RAN architecture makes it hard to satisfy all requirements adopted by 5G. As a result, an innovation of RAN architecture with high capability, flexibility, and scalability is expected to free this gap, which leads to the adoption of CU/DU architecture.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

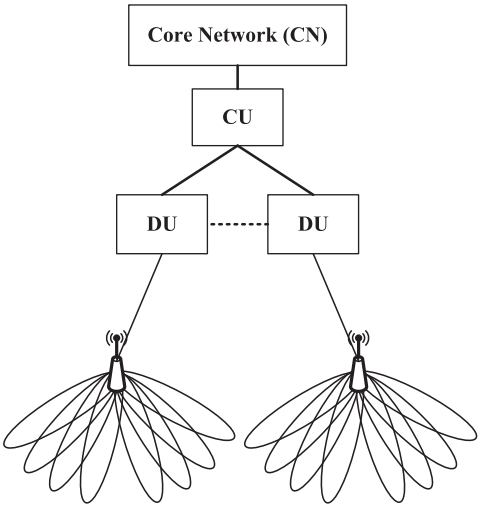

Figure 3.7 CU/DU architecture.

Figure 3.7 shows a conceptual CU/DU architecture with one CU and several DUs. In a CU/DU architecture, the CU can support multiple DUs and each DU belongs to only one CU. In general, CU is responsible for data processing as well as controlling the context of users, the profiles of services, and the connections between the DUs. The DU is responsible for scheduling radio resources and physical layer numerologies.

More specifically, CU controlling could be classified into the intercell level and user level. For intercell-level controlling, the CU keeps track of the capacity per cell and informs the DU to perform interference coordination, load balancing and power control among cells that ensure the wireless coverage area (note that each DU is in charge of at least one cell, and each cell only belongs to one DU). For the user level controlling, the CU selects a suitable DU for UE based on its channel quality and DU capacity. By CU scheduling, no data forwarding is needed within a single CU, which is advantageous in Ultradense Network (UDN) scenarios.

DU scheduling is based on signal quality, user mobility, radio link robustness, and air interface numerologies. Under the CU controlling, the DU should anchor to a CU and be mastered by the same. In addition, the DUs might be logically interconnected with each other. The functions fulfilled on the DU can be split into two parts, which are implemented by two functional modules. The functional module for radio resources scheduling performs allocations of control channels, resource blocks (RBs), and power, achieving the signaling procedures and timing relationship over the air interface. The functional module for physical layer numerologies fulfills the fundamental processing operations, such as modulation, multiplexing, channel coding, and channel measurement. Besides, the DU should report status information to the CU and cooperate with the CU to fulfill the maintenance of the CU/DU architecture, which is easy to be extended according to performance provisions.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

A 5G base station, called a gNodeB (gNB), is comprised of the CU, the DU, and the RRU. Compared to LTE eNodeB (eNB), which only contains the BBU and the RRU, new RAN architecture further divides the BBU into a CU and one or multiple DU(s). Most of the controlling functionalities are centralized on the CU while the fast scheduling on the air interface is realized on the DU. In fact, different functions for the CU and the DU reflect the more precise division of responsibilities than the original BBU, in order to achieve high capability, flexibility, and scalability. The CU is able to pile up the resources that enable the manipulation of a number of DUs, and the DU is able to handle multiple numerologies that indicate different processing requirements; thus, the high capacity can be achieved. Furthermore, the CU and the DU can be deployed independently according to the requirements of services and the capacities of the hardware; thus, high flexibility can be achieved. Thirdly, the CU is suggested to use general hardware, which is easy to be extended, and the number of the DUs connected to one specific CU depends on which is on demand and dependent on the capacities of the CU and the DU. Thus, high scalability can be achieved.

Feasible Functional Split Options

The functional split of the protocol stack is the key to CU/DU architecture design. In the study item for a new radio (NR), 3GPP RAN3 has investigated several possible functional splits based on E-UTRA protocol stack. As shown in Fig. 3.8 [12], a total of eight split options have been identified.

Roughly speaking, the more parts of protocol stack reside in the CU, the higher pooling gain the system can achieve. However, this situation may introduce the transmission latency of the interface between the CU and the DU. As a result, the judgment of the feasible functional split should concentrate on data processing capability, implementation complexity, and transmission latency.

3GPP RAN3 has divided the options into two categories, i.e., higher layer split and lower-layer split. More specifically, options from 1 to 3 are classified as higher-layer split while options from 6 to 8 are lower-layer split. Option 4 is not taken into consideration since its benefit is unforseen. In addition, the classification of Option 5 is controversial and will be investigated independently.

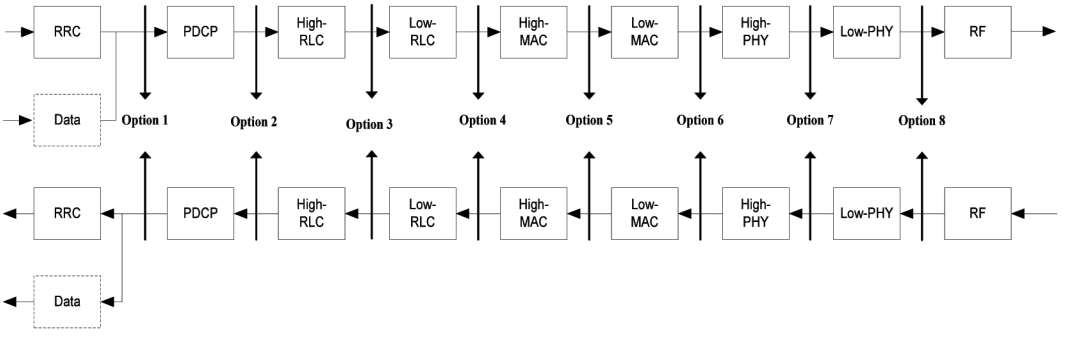

Figure 3.8 Functional split between CU/DU [12].

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

Based on the current process of 3GPP RAN3, Option 2 (PDCP/RLC split) is suggested as the normative higher-layer split after a down selection between option 2 and 3–1 (intra-RLC split, as specified by 3GPP). The reasons to standardize option 2 are listed as follows:

1. Option 2 adopts the standardized LTE dual connectivity (DC) split as the baseline, which requires relatively small normative work compared to option 3–1.
2. The ARQ at the RLC layer of option 2 is located on the DU, which achieves lower retransmission latency compared to option 3–1, for which the ARQ resides in the CU.
3. With certain enhancement, option 2 can achieve comparatively fast centralized retransmission of lost RLC PDUs compared to option 3–1.

However, for lower-layer split, further study is needed to justify the benefits of each option. Although the determination of lower-layer split is still an open issue, several common views can be summarized as follows:

- Generally speaking, apart from option 8, which has been used in LTE and will not be detailed in new radio (NR), all the other lower layer split options suffer from the transmission latency between the CU and the DU. Therefore, in our opinion, an ideal FH condition should be premised. Otherwise, the lower layer split brings no benefit.
- One of the most important advantages for the lower-layer split is the realization of centralized scheduling, which is a key function to achieve the intercell level controlling as specified earlier.

In summary, the determination of the functional split should be based on the following two aspects. On the one hand, most of the controlling and centralized scheduling functions are suggested to be located on the CU while the fast scheduling functions are suggested to be located on the DU. On the other hand, the difficulty of implementation and standardization should be kept within a scope while ensuring the capability for future optimization.

Interfaces

After determining functional split, the interfaces should be defined. The interfaces related to the CU/DU architecture can be classified into two categories. The RAN-core network (CN) interfaces, and the interfaces between the CU and the DU. For RAN-CN interfaces, the CU, DU, and RRU can be treated as one logical node, namely gNB, which is regarded as the termination point of the NG interfaces. Based on the current agreements of 3GPP RAN3, the standardization of the NG interfaces sets LTE RAN-CN interfaces, i.e., S1 interfaces, as a baseline. Therefore, the NG interfaces won't be detailed in this section. On the other hand, the interfaces between the CU and DU, or F1 interfaces, may adopt the same protocol stack as in LTE for both the user plane and the control plane, since the legacy protocol stack has proved to be sufficiently reliable. However, because of the introduction of functional split and the characteristics of the CU and the DU, more F1 application protocol (F1AP) should be added in order to ensure

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

the normal operations and possible optimizations. Although the detailed descriptions of F1AP are still under discussion, in our opinion, there are two new aspects that need to be considered:

- Since radio resource control (RRC) is operated on the CU, the F1 interface should support the transfer of the RRC messages between the CU and the DU.
- Since the CU is responsible for managing multiple DUs, the F1 interface should consider including management functions to support operations, such as flow control and load balancing among DUs.

Note that the normative work of CU/DU interfaces is expected to be done in 3GPP's specification TS 38.401. We suggest that the reader directly refer to TS 38.401 Release 15 for more details.

Mobility

The implementation of the CU/DU architecture enables RAN to optimize functions of the protocol stack and procedures of operations. One of the most vital advantages of the CU/DU architecture over 4G is mobility. For example, seamless switch, which is one of the objectives that 5G RAN must meet to satisfy requirements of services with low latency, can be realized by means of CU/DU implementation. By adopting the CU/DU architecture, the radio link of UE can be split into the CU-level part and DU-level part.

The CU-level radio link is mainly in charge of the data processing of UE, which means two types of basic information are needed: the context of the radio link and the data transmitted on the radio link. The context identifies a radio link including identifiers of RBs and logical channels, formats of PDUs and SDUs, and the mapping between RBs, logical channels, and transport channels. With the aid of the context of the radio link, the data can be accurately transmitted and received.

The DU-level radio link is mainly in charge of the transmission and reception of transport blocks (TBs) over the air interface, which means two types of basic functions are needed: the function of selecting an available physical channel and the function of assembling suitable TBs. The scheduler performs the selection and assembling, and matches TBs and physical channels dynamically.

Given the difference between CU-level and DU-level radio link, the inter-CU and intra-CU mobility management of UEs should be treated separately.

When the inter-CU switch is triggered, both CU-level and DU-level radio links should be switched. For the CU-level radio link, data forwarding is required, and the context of UE should be forwarded to the target CU in order to achieve lossless switch. For the DU-level radio link, the seamless switch cannot be achieved because both the CU-level and the DU-level radio links need to be reestablished. However, by switching the CU-level and the DU-level radio links separately, the interruption time over air interface could be minimized. First, a cloned CU-level radio link can be established on the target CU while maintaining the whole radio link on the source node. Second, the DU-level radio link is quickly switched from the source DU to the target DU by CU scheduling. Note that it takes a longer time for the CU-level switch because of CU-level data forwarding.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

When the intra-CU switch is triggered, no data forwarding is required and no switching latency is caused. Therefore, the seamless switch of CU-level radio link can be achieved. For the DU-level radio link, the DU can fulfill seamless switch if the DU schedules new users and resource blocks, which are TTI-based. More specifically, if the switch of a partial radio link can be scheduled within one TTI, the seamless switch is achieved within a DU; while the seamlessness is nearly achieved among DUs if the switch of a radio link can be scheduled within several TTIs. The value of TTI can be dynamically configured by the scheduler according to each radio link switch.

In summary, the partition of the CU-level and the DU-level radio links is expected to simplify the switch procedure and reduce the switch latency, especially for the intra-CU switch scenario. Intra-CU switch achieves seamlessness by integrating radio link switch into scheduling operation, which is a significant advantage of the CU/DU architecture.

3.2.6 Rethink Protocol Stack for 5G: MCD

Motivation

ITU-R has defined three key scenarios for 5G including eMBB (10 Gbits/s), URLLC (99.999%, 1 ms) and mMTC (1 million/km^2), each of which provisions stringent requirements in different aspects. From the aspect of RAN, the classical 4G RAN is not capable of satisfying all of these demands simultaneously. For massive data scenarios and the deployment of more dense nodes in 5G, multi-RAT for physical layer (PHY) and BD computing capability based on cloud platforms are proposed. As a generalization, five innovative R&D themes have been proposed in [32] for 5G RAN. As indicated in [32], with the introduction of UDN in 5G, the importance of flexible air interface is highlighted, which is especially reflected by rethinking of Ring and Young, which represents an effort to review traditional cell-centric network design and put forward user-centric design by adopting the concept of "no more cells" (NMC). With NMC, the available radio resources from multiple access points could be jointly scheduled dynamically for each user and the selection of control/user plane and uplink/downlink (UL/DL) channels respectively could be done separately. With the development of the 5G RAN architecture, the emergence and the adoption of C-RAN by the industry has facilitated the realization of the concept of NMC.

From the perspective of the protocol stack, the signaling interaction in protocol stack architecture is complex, although the framework of the traditional LTE protocol stack architecture is clear. More specifically, in traditional LTE/LTE-A, the basic element of communication network is "cell" that manages the radio resources and the users connected with it. As shown in Fig. 3.9 in the traditional LTE protocol, the UE context can only be established based on a specific cell. Even in carrier aggregation (CA) scenarios, the UE context is established based on the primary cell (PCell) rather than secondary cell (SCell). The SCell only provides channels for data transmission/receiving. In the procedure of cell handover, signaling interaction between cells is slow, and the duration is on the order of seconds or even minutes, which is in the same case as the signaling interaction in some technologies, such as intercell interference coordination (ICIC).

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

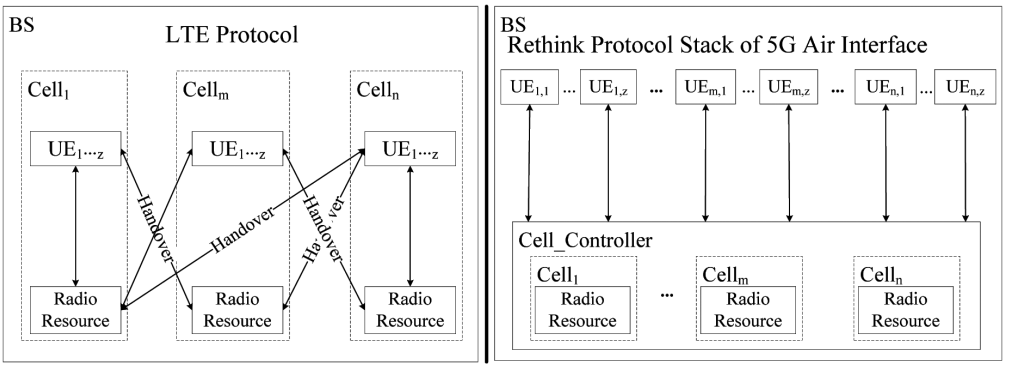

Figure 3.9 Rethink protocol stack.

In 5G, a user-centric network (UCN) is introduced [44] to solve the problem of explosive growth of data traffic and increasing density of deployed BSs. The signaling procedure in the UCN reflects the concept of CP/UP decoupling. According to the quality of the air interface, the network should provide corresponding radio services in order to maintain the connection of CP/UP and the transmission of signaling and data [45]. To improve the quality of the air interface, many new air interface technologies have been introduced to 5G networks, e.g., full duplex, hybrid beamforming PHY, and so on. However, those new technologies bring new challenges to the 5G air interface, requiring the network to provide corresponding services to UE to meet the demand for data rate and channel quality in every TTI. In order to support the requirement, the coordination among cells should keep real-time in each TTI, which greatly increases the difficulty of processing on the network side.

As a result, it is necessary to rethink the protocol stack of air interface for the requirements of 5G and the status of the traditional networks. The traditional network, which is characterized as cell-centric, has been proved to be a simple and practical method of radio resource management [46]. The protocol stack of the 5G air interface should inherit the advantages of the traditional networks, which will be reconsidered to meet the requirement of 5G network and coexist with the traditional network protocol stack.

In summary, the traditional LTE protocol stack architecture is not suitable to support the UDN scenario because of the signaling overhead caused by frequent handover, which motivates us to rethink the protocol stack for 5G air interface. Following the concept of NMC, the protocol stack architecture should be "user-centric" to provide flexible air interface and reduce frequent radio resource control (RRC) signaling transmission. Meanwhile, the innovative protocol stack should take full advantage of "cloud" with strong computing capability. Considering the high density of users and cells, the protocol stack architecture should implement the optimized configuration of air interface resources, including frequency domain resources, time domain resources, and air space resources. As an outcome of rethinking the protocol stack, MCD design logic has been proposed, aiming at achieving the goal of NMC for 5G NR.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

The MCD protocol stack is a unified and seamless mobility-enabling framework, which is able to fulfill the same functions as other multicell frameworks adopted by 3GPP, including CA [45] and dual connectivity (DC) [46], with the help of the innovative modification on MAC scheduling and RRC signaling. First, the MCD protocol stack achieves the seamless handover by means of MAC scheduling, which is different from the CA case. Furthermore, the MCD protocol stack achieves the seamless handover without the establishment of an additional RLC entity, which is different from the DC case. Consequently, the MCD protocol stack reflects the concepts of green and soft. On one hand, the data forwarding procedure during the handover within the same BS, is not required, which saves the energy of the BS and the idea of green is achieved. On the other hand, the functions of each layer can be dynamically adjusted by different service requirements, which embodies the idea of soft.

More Details in Differences

According to the analysis just discussed, we have proposed the concept of MCD protocol stack of 5G air interface, in which the UE makes decision independently. In our proposal, both "UE" and "cell" are the basic elements of communication networks, and cells become the dedicated radio resource management elements.

The difference between the protocol stack "rethink" and the traditional LTE protocol stack is shown in Fig. 3.9. In the traditional LTE protocol stack, all the signaling and context of UE treat the connected cell as the only key label. For example, the scope of cell radio network temporary identifier (C-RNTI) for each cell is 0-65535 [47]. The data radio bearer (DRB), signaling radio bearer (SRB) (and the mapping process of SRB/DRB to E-UTRAN Radio Access Bearer (E-RAB) of UE) are both allocated and managed in the scope of one cell. With CA, UE can use the resource of more than one cell. However, as a supplement technology of the LTE protocol stack, CA cannot assist the LTE network to solve the problem in 5G since CA is still cell-centric. Generally speaking, the adoption of the cell-centric scheme is a double-edged sword. On one hand, with "cell" as the key label, the LTE protocol stack simplifies the process of radio resource management for the network [48]. On the other hand, such protocol stack increases the complexity of management and leads to high delay in UE mobility, which is hard to fit into the needs of 5G.

In the MCD protocol stack, "UE" is also a basic element as well as "cell." On one hand, as the element of the protocol stack, UE is responsible for the management of all its information, including its context, the mapping process from DRB/SRB to E-RAB, its channel quality, and the dedicated radio resources allocated, etc. On the other hand, as the only element of the original protocol stack, the cell manages all the radio resources that are not allocated to any users, including ICIC radio resources [49]. As shown in Fig. 3.9, the Cell_Controller allocates the radio resources from different cells to fulfill the requirement of a specific UE according its specific requirements. The radio resources allocated to the UE become its specific attribute, and the UE returns radio resources to cells when the transmission process ends [50]. Consequently, the resource allocation and release are just related to changes of the UE attribute, which

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

works in the same way as the UE context setup/modification. Such an operation avoids the changes of DRBs and logical channels when radio resources change. As a result, the procedure of handover is replaced by the modification of UE attributes for radio resources, which is expected to be carried out in a much faster way than in the traditional structure. In addition, the modifications of semi-static links only occur when the inter-gNB handover is required, while for the intra-gNB inter-cell switch, no modification is needed, which dramatically reduces the frequency of necessary handover, especially for the scenarios with high-density cells.

The MCD Protocol Stack

The functional blocks of a cell and UE in the MCD protocol stack for 5G air interface is shown in Fig. 3.10. The functions of a cell include the following two parts. The Cell_Signaling Controller module is in charge of the air interface signaling while the Cell_UU (UE to UTRAN) Controller module is responsible for the radio resource allocation. The functions of UE include the following two parts. The UE_Packet Processing module is in charge of the data packet processing while the UE_Channel Measurement module is responsible for the collection and analysis of the measurement results [51, 52].

The Cell_Signaling Controller module corresponds to the traditional RRC protocol in the LTE protocol stack [53], which manages the radio resource at a time interval much

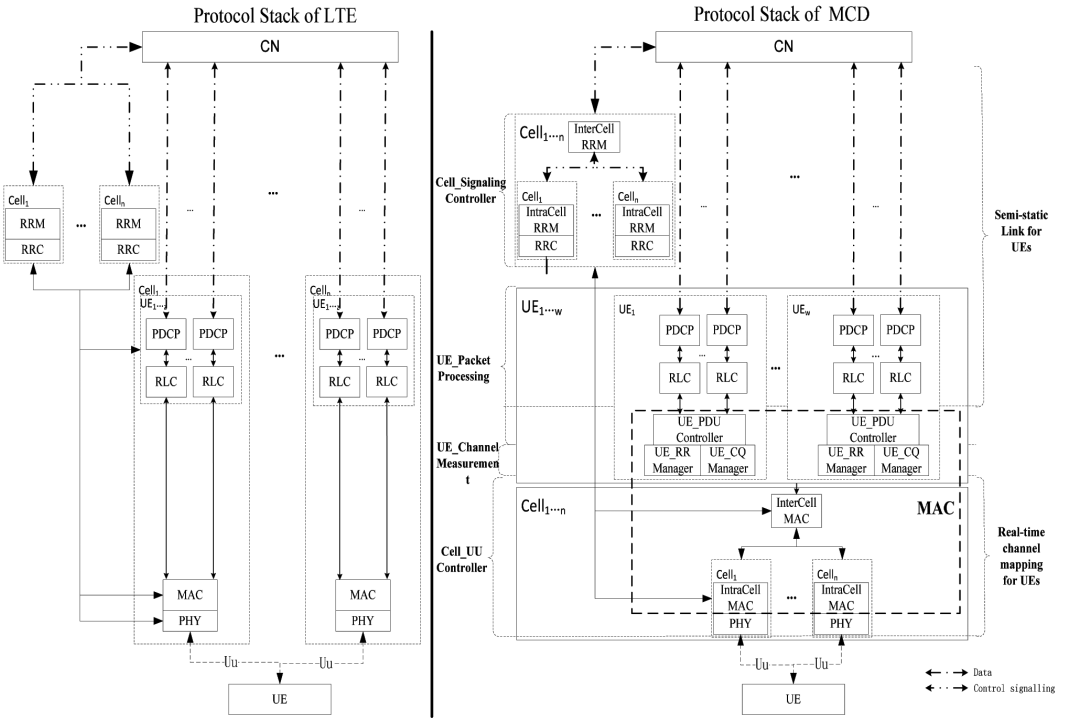

Figure 3.10 "Independent UE" protocol stack.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

larger than TTI. More specifically, the Cell_Signaling Controller module is comprised of the RRC module and the radio resource management RRM) module. In the MCD protocol stack, the RRM module is further divided into InterCell RRM and IntraCell RRM.

The Cell_UU Controller module corresponds to the PHY protocol [54] and most part of the MAC protocol in the traditional protocol stack. It fulfils the radio resource allocation between InterCell MAC and IntraCell MAC and the mapping process from UE data to the physical channel. It should be noted that the division of the original RRM and MAC modules into intercell and intracell parts is a reflection of the sinking of functions related to the cell.

The UE_Packet Processing module corresponds to the original PDCP/RLC protocol in LTE, and the UE_PDU Controller. The PDCP/RLC module achieves the data packet processing and provides the service access point (SAP) of the UE context while the UE_PDU Controller module controls the collection and dissemination of PDUs for concurrent PDU streams [55, 56].

The UE_Channel Measurement module corresponds to the channel control part of the MAC protocol, which includes the UE_CQ (channel quality) Manager and the UE_RR (radio resource) Manager modules. The UE_CQ Manager module is in charge of monitoring, modification, and computation of the channel quality, which provides the information to support the MAC scheduling. The UE_RR Manager module is responsible for the management of UE specific radio resources.

To meet the stringent requirements of three typical scenarios adopted by 5G, the MCD protocol stack provides an innovative pattern for the link control, which is the collaborative management of semi-static links and real-time channel mappings.

The semi-static links for UEs is comprised of logical channels, DRB/SRB, and E-RAB links, all of which indicate a specific type of service (ToS) of the UE and work in a semi-static way. To achieve more flexibility, it is necessary to unbundle the ToS with a specific PHY mode, which decouples the UE from the cell. As a result, the semi-static links only need to be modified when inter-BS handover occurs.

The real-time channel mapping for UEs is responsible for the mapping of logical channels, transport channels, and physical channels. The UE provides parameters to MAC including the quality of channels, the buffer state, the request to PHY, and the characteristics of allocated radio resources. According to radio resources of all the available cells and the parameters received, MAC configures cells together with their radio resources as attributes of the UE. By this means, the logical channel matches the appropriate cells at first, and then performs channel mappings within each cell. With real-time channel mappings, UE can receive data from different cells.

With the combination of semi-static links and real-time channel mappings for the UE, the MCD protocol stack optimizes the handover procedure compared to the traditional LTE protocol stack. As shown in Fig. 3.11, in traditional LTE protocol stack architecture, handover is required as soon as the UE moves across cells. In comparison, in the MCD protocol stack architecture, the real-time mapping replaces the handover within the same BS, which enhances the flexibility of the radio resource management

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

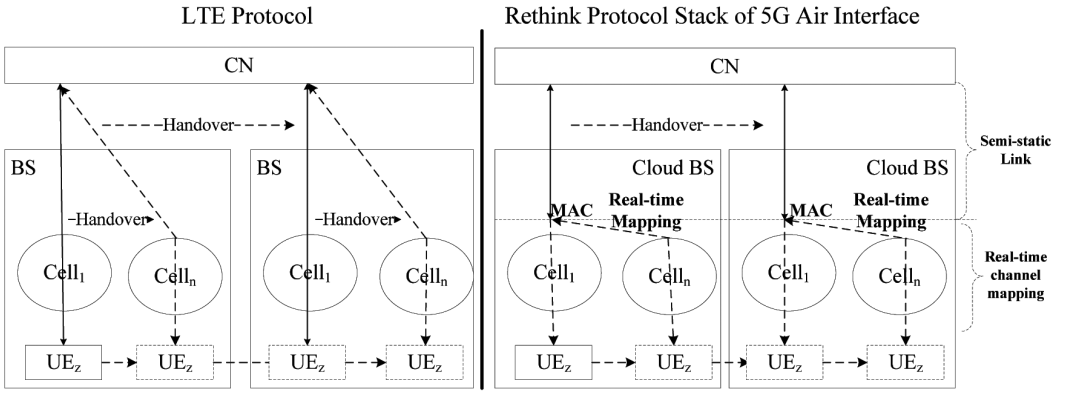

Figure 3.11 The mobility of UE.

while ensuring the stability of the system by redefining the relationship between the cell and the UE.

The Application to C-RAN

As mentioned earlier, the concept of C-RAN has been widely accepted by the industry for 5G. One of the most important characteristics of C-RAN is the adoption of the cloud platform, which in theory provides sufficient hardware processing capabilities for the management of the air interface. However, in reality, the hardware resource of the cloud platform is relatively limited, and so is the network scale of 5G wireless network; thus, the mapping between the hardware resource and the network scale needs more research.

The cloud platform is able to achieve the strong computing capability and dynamically adaptive hardware management capability. The MCD protocol stack can make the most of the cloud platform in terms of intercell management for the performance optimization.

As indicated previously in the section, the MCD protocol stack redesigns the architecture of the protocol stack into the centralized and distributed levels, which achieves the centralized link control and the real-time collaboration of the air interface across cells. On one hand, the centralized functions can make full use of the strong computing capability of the cloud platform. On the other hand, the distributed functions make dynamical adjustments according to the load of the network, which makes full use of the adaptation ability of the hardware resource management of the cloud platform.

A possible design for the MCD protocol stack with the introduction of the cloud platform is shown in Fig. 3.12. From Fig. 3.12, the MCD protocol stack isolates the PHY and IntraCell MAC from InterCell MAC, and only the InterCell MAC is implemented on the cloud platform. As a comparison, PHY and IntraCell MAC are still implemented on the dedicated hardware devices. As a result, the original functions of MAC can be split into two parts: intercell control for UEs (all the other blocks in the MAC module except IntraCell MAC in Fig. 3.12) and IntraCell MAC. On one hand, InterCell control for UEs realizes seamless handover and performs operations including flow control

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

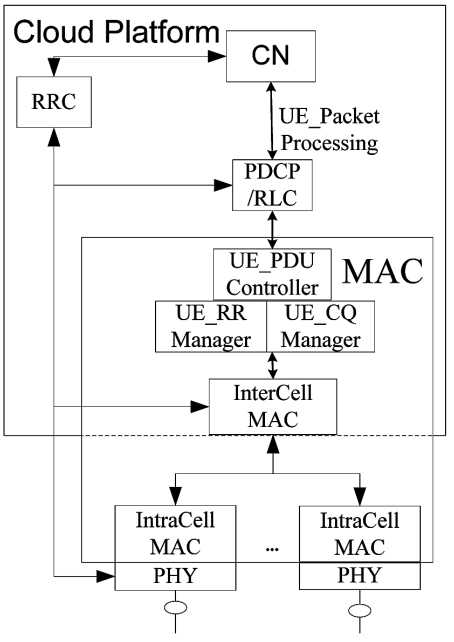

Figure 3.12 The cloud platform for the MCD protocol stack.

and load balancing across cells. On the other hand, IntraCell MAC schedules UEs within a cell and performs operations, including transport/physical channel selection, service mapping, and priority control, which are under strict TTI constraint over the air interface.

The Progress of 5G NR Protocol Stack in 3GPP

The progress of 5G NR protocol stack is mainly obtained in [57]. According to discussions and agreements in 3GPP, NR protocol stack has been enhanced in the following aspects:

- New AS layer
  In NR, a new AS layer is added on top of the PDCP layer that is in charge of the mapping from QoS flow to DRB. The new AS layer is called the service data adaptation protocol (SDAP) layer. Originally in LTE, there is only one-to-one mapping between E-RAB and DRB, and each E-RAB only corresponds to one set of QoS parameters; while in NR, E-RAB is replaced by QoS flows, and multiple QoS flows can map to the same DRB. As a result, 5G NR can acquire the characteristics of service streams more precisely compared to LTE. In addition, the SDAP entity is configured per PDU session, which reflects the idea of user-centric design logic.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

- PDCP
  As mentioned in the previous section, according to the current progress of RAN3, the functional split option 2 has been adopted as the normative split for the CU/DU architecture, which splits between PDCP and RLC. As a result, the functions that are suitable to be operated on the cloud platform are centralized to the PDCP layer. Therefore, in 5G NR, the reordering function has been moved from RLC to PDCP. And the duplication function, which is newly introduced to improve the reliability of services in NR, is added at the PDCP layer. Moreover, the dual connectivity adopted by NR has set PDCP as the anchor of the data split bearers, which results in the introduction of other functions, including data dispense and flow control at the PDCP layer.
- RLC
  As indicated by NR, one of the main objectives of the 5G protocol stack is to minimize the processing delay at the RLC layer. Consequently, the reordering function is moved from RLC to PDCP, and the concatenation function at RLC is merged to the multiplexing function at MAC. In addition, in order to further optimize the processing delay, the concept of preprocessing is introduced to both RLC and MAC layers.
- MAC
  Since the concept of the numerology is introduced at PHY, 5G MAC is enhanced accordingly compared to LTE. Besides the TTI length, the mapping relationship between logical channel group (LCG) and the numerology should be reconsidered, and Buffer status report (BSR), scheduling request (SR), logical channel prioritization (LCP), and discontinuous reception (DRX) in NR will also be impacted. In addition, from the perspective of random access (RA), multiple beams will support different set of RA parameters, such as backoff and power ramping. Moreover, in order to achieve the fast activation of the duplication function at the PDCP layer, a new MAC control element (CE) is introduced at the MAC layer.

  The detailed functions for L2 layers in NR are shown in Fig. 3.13. Note that the main enhancement for the protocol stack is marked with circles. In summary, all these modifications and enhancements reflect the thought of decoupling UE from cell. Combined with the adopted CU/DU architecture by 3GPP, CU fulfills the more UE-specific centralized control, while DU achieves fast scheduling functions that are more cell-related. Therefore, the current progress of 3GPP on the protocol stack matches the MCD design logic proposed in this book.

Besides, in order to further decouple UEs from cells, 3GPP has introduced the concept of beam at the air interface. Beams are mainly operated at the MAC and the PHY layers. A specific UE can be served by a dedicated beam during a time period based on the beam tracking, and the fast beam switch can be achieved at TTI level. In addition, the division of traditional RRM is under discussion in 3GPP RAN3. As indicated by the latest progress of RAN3, 5G NR architecture is striving to sink part of RRM functions, which are more cell-related, down to DU. For example, 3GPP is discussing to achieve C-RNTI

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

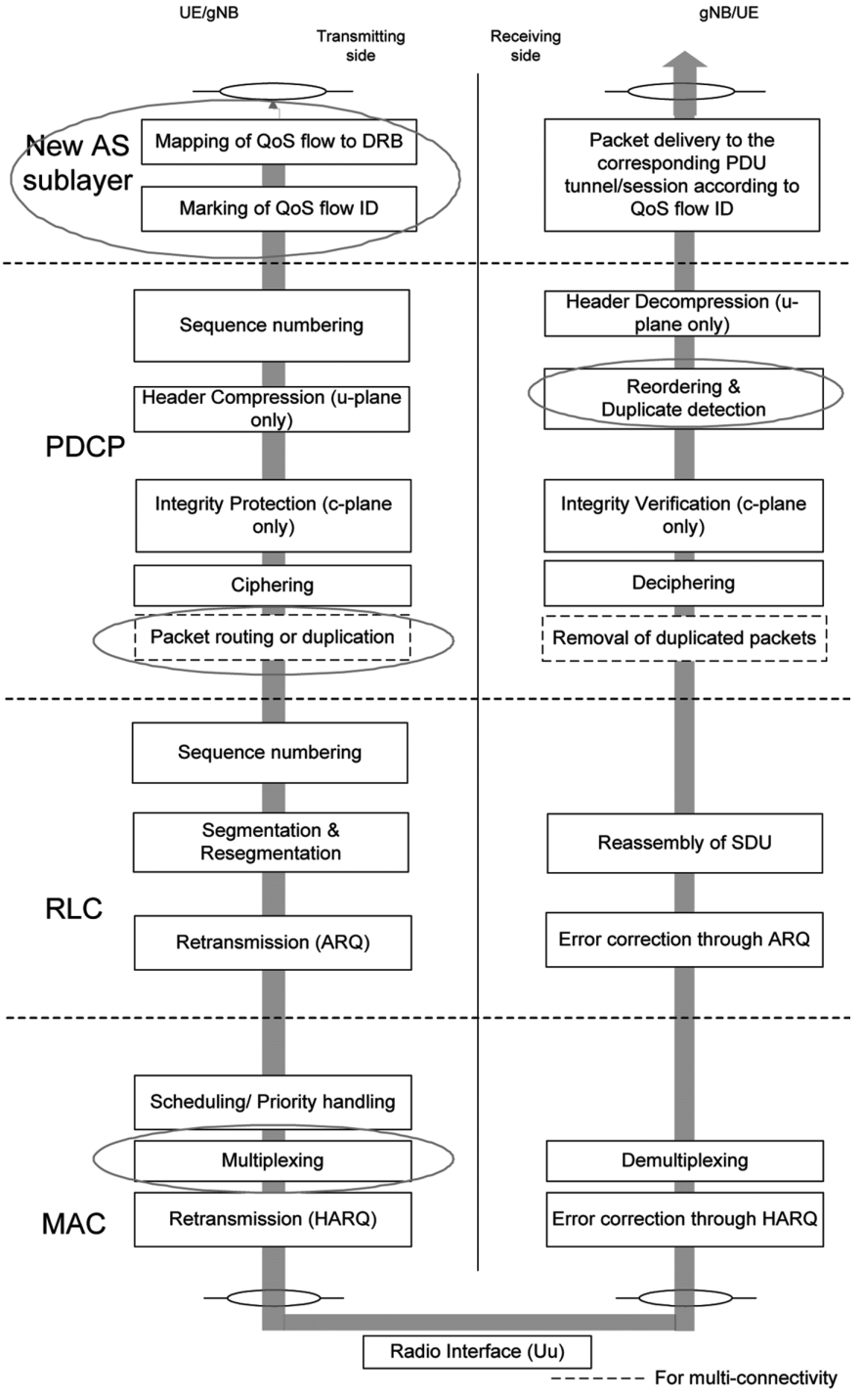

Figure 3.13 The 5G NR protocol stack [57].

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

allocation in DU since C-RNTI is a cell-specific parameter. As a generalization, the beam management and the sinking of RRM functions again reflect the idea of the MCD design logic.

3.3 Big-Data-Enabled Mobile Network Design

There is a broad agreement about the remarkable potential of big data to lead innovation, boost commerce, and drive progress, among leaders in the industry, academia, and government. The term "dig data" can be literally understood as the surge of data in today's networked, digitized, and information-driven world. With vast data resources, people can tackle questions previously out of reach. Additionally, an agreement is also reached on the inevitable fact that BD will overwhelm traditional analytical approaches, architecture, information management, and data transport. In this chapter, we begin with the consensus on some important basic issues with BD, such as definition, characteristics, history, and mainstream technology. Then, it is followed by a special focus discussion on wireless BD and its applications into the mobile network, seen from a network operator's perspective. And last but not the least, we present the BD-driven mobile network design, as well as some initial investigative results.

3.3.1 Background of Big Data

The rate of growth of data generated and stored has been increasing exponentially. The 1965-born Moore's Law has been applied to almost every aspect of the computer industry, from integrated circuits to memory, whereas the growth rates of data volumes are estimated to be faster than Moore's Law, i.e., more than doubling every eighteen months.

This data explosion sparks new ways of gathering and utilizing data to extract value, meanwhile providing significant challenges due to the size of the data being manipulated and studied. Another significant shift stems from the increasing amount of unstructured data. Historically, enterprise data analytics has been focusing on structured data and using relational data models to capture data characteristics, and discarding the fragmented and unstructured "noise." With the soaring of the quantity of unstructured data, such as web pages, texts, images, and videos, there now is a strong desire to obtain additional value from this heterogeneity nature of today's data. The ability to process not only a large amount of data but also various types of data leads to the current revolution to parallel scalability in the architecture and subsequent data handling methods.

To understand this BD revolution, the following four aspects should be considered: the characteristics of the datasets, the historical milestone events, the relation between internet technology and data technology, and the state-of-the-art scientific and technological tools. The following sections will briefly discuss these aspects.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

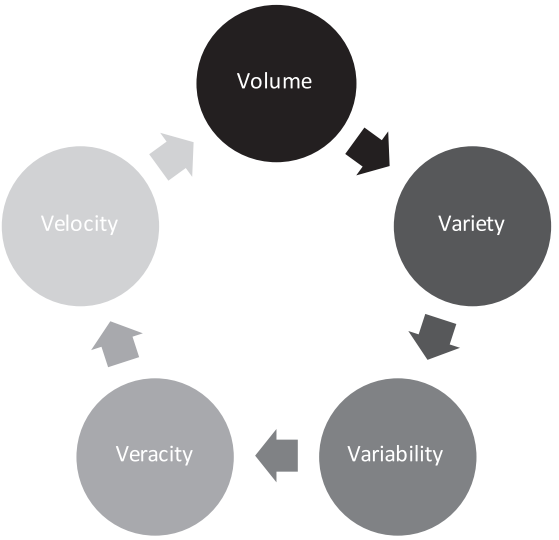

Figure 3.14 5V characteristics of big data.

Understanding Big Data

BD is a broad concept where different industries have different understandings and thus different definitions [58]. Nevertheless, BD can be distinguished by the following five characteristics as in Fig. 3.14 :

- **Volume**, the most commonly recognizable feature, i.e., the presence of the extensive data sets available for extraction of value. The underlying assumption here is that the larger the volume the greater the value.
- **Velocity**, a contextual reference to the speed at which the data is generated and processed, often in real-time, or near real-time. It is opposite to the data at rest.
- **Variety**, the need to process data from multiple repositories, types and domains and the need for analytics across a range of logical models, timescales, and semantics.
- **Variability**, the changes in a dataset's characteristics; whether it is infrastructure, interface, flow rate, or volume, all have impacts on data processing (whereas volatility refers to the changing values of the actual data elements in a data set).
- **Veracity**, the varying quality and creditability of the data, which comes from multi-sourced and multidimensional inputs and affects the accuracy of the analysis.

To reference the National Institute of Standards and Technology (NIST) Big Data Working Group (NBD-WG) in the United States, it describes BD as the datasets that are so large or complex that even some current data processing application software and data analytics cannot adequately deal with. Scientific and technological advances have been outrun by the rapid growth of data. Therefore, innovative technical approaches in

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

information processing are needed to reap the fruits of BD, such as enhanced insight, decision-making, and process automation.

Historical Journey

The BD industry has gone through roughly three main development stages:

1. 2003–2006: BD development was still in its infancy. The explosion of unstructured datasets such as text and video that do not fit into predefined relational data models, and dataset research gave rise to scientific and technological advances. The milestone event was that Google published three influential papers that introduced its Google File System (GFS), MapReduce, and BigTable to the world. All three core technologies are distributed in nature, and able to process huge amounts of unstructured data using cost-effective hardware and software implementations. The most notable application was search engines.
2. 2006–2009: BD saw some major breakthroughs. Parallel computing and distributed processing became the pillars to BD's fast-paced advances. With reference to Google's technical architecture, Hadoop, an open-source software framework used for distributed storage and processing, emerged and was soon used by IT companies to build data processing systems. By leveraging users' web browsing behavioral data, IT companies were generating user profiles that in turn guided targeted advertising and promotions.
3. 2010–now: With the widespread adoption of mobile phones, mobile data usage skyrocketed. BD development has entered a new era, when the data is more fragmented, distributed, and optimzed for streaming media than ever before. There are many novel BD analytics tools that have originated from the open-source community, for instance, Storm, S4, and Spark. Innovative applications also flourished, ranging from real-time marketing and internet credit investigation to vertical application solutions for industries such as tourism, real estate, transportation, and business.

IT vs. DT

Information technology is all about data processing and manipulation, and it provides the software tools for people to use predefined procedures to conduct business. The analogy is that IT has extended the reach of human beings' arms and legs. Data technology focuses on data analytics and provides the means to extract information and produce new knowledge. The analogy is that DT is expanding human beings' brains. Therefore, the difference here is not only in the methodology, but it is rather a paradigm shift.

It can be predicted that there is a revolution in the near future where data defines and determines the technological progression, resource distribution, fund flow, and talent flow in society. It will fundamentally shake traditional social structure and cooperation, leading to more effective and efficient modes of production and economic operation. The IT industry's motto is to own, to manipulate, to transport, and to control; DT industry's slogan is rather to open, to interwork, to experience, and to share. DT is taking over IT as the fundamental driver of social progress.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

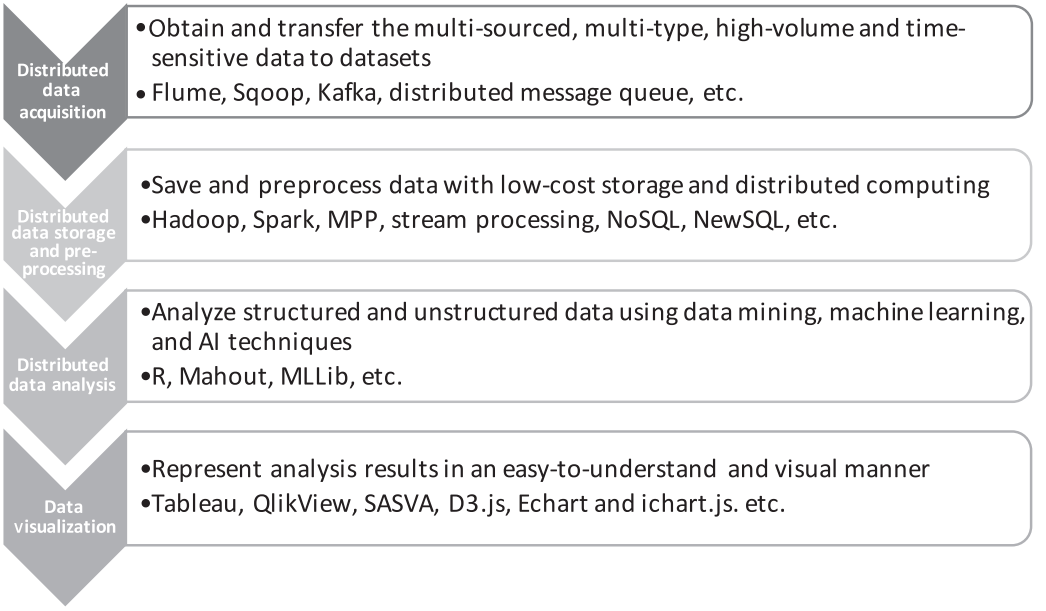

Figure 3.15 Big data life cycle and main analytical tools.

Mainstream Technology

In order to meet the challenging requirements of BD applications and massive datasets, the technical community has come out with numerous data analytics and solutions. From the perspective of the life cycle of data in the system, there are four processes: data acquisition, data storage and computing, data analysis, and data visualization, as shown in Fig. 3.15.

1. **Distributed data acquisition**
   This process includes obtaining the source data and transferring them to datasets. Due to the multi-sourced, multi-type, high-volume, and time-sensitive features of the captured data elements, it is critical to have cloud computing and a distributed system architecture. The mainstream techniques are Flume, Sqoop, Kafka, and distributed message queue.
2. **Distributed data storage and computing**
   The amount of data is growing at such a rapid rate that it demands extremely low-cost storage capability to handle the diverse forms of data, as well as the massive computing capability to preprocess the raw data. Traditional CPU and parallel computing can no long meet the requirements on speed, scalability, and cost. Distributed computing becomes the dominant method, including Hadoop, Spark, MPP, stream processing, NoSQL, NewSQL, etc.
3. **Distributed data mining**
   Data-mining techniques are used to discover new knowledge and new relations from the chaos of data, which is the primary focus of any BD system. Traditional

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

data mining was performed on structured and small datasets, where predefined experience-based relational models were established and then used in data analysis. However, little prior knowledge is known about unstructured and multi-type data, let alone used for building an explicit mathematical model. More intelligent and distributed non-relational data platforms should be investigated. Some of the analytic tools are R, Mahout, MLLib, etc.

4. **Data visualization**
   Representing the data analysis results to the user in a straightforward and meaningful manner is seen as a crucial segment in decision-making scenarios. The biggest challenge is how to make the complex outcome easy to understand, and there are a few current techniques, including Tableau, QlikView, SASVA, D3.js, Echart, ichart.js. etc.

3.3.2 Wireless Big Data

The global success of the wireless industry is undeniable, namely cellular networks, internet of things (IoT), Wi-Fi, satellite, and sensor networks, as well as many private and dedicated communication networks, such as smart grid and intelligent transportation networks, that are traditionally wired communication networks but are migrating to the wireless domain.

The explosion of wireless data and the increasing number of types of data has demonstrated that the wireless communication industry has entered the age of BD. Wireless big data (WBD) applications are borne out by the realization that there is an enormous amount of system data generated each day (estimated more than 14 Petabytes/day from a dozens of CMCC data centers alone) and should be properly analyzed and used for improving the ecosystem comprised of users and the network.

- **Definition, Sources, and Classification**
  There is no universally agreed definition of WBD as of today, yet it is generally described as the massive data consumed by wireless service users, and the data generated by the wireless system when providing such services. This may relate to wireless spectrum, transmission, access, and network, etc.

  With respect to the sources of the WBD, we initially proposed three categories: raw WBD, derived WBD, and trial WBD. Raw WBD is the data, old and new, that has passed through wireless communication systems, originating from content providers, end users, and the network itself. Derived WBD is the statistical analysis that is beneficial to service delivery, such as spectrum utilization distribution, small-cell deployment sites statistics, and radio resource allocation statistics. Trail WBD is the acquired performance data for any new trial and testing effort, either for new transmission techniques or innovative functional and architectural proof points. Alternatively, WBD can be divided into four categories according to [59], as shown in Table 3.1, i.e., application data, user data, network data and link data. The rationale is explained in details for each category with the means of acquisition. The application data usually entails the statistics of content popu-

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

Table 3.1 Four categories of wireless big data

Data Category	Contents	Big data vs. Traditional data
Application data	Content popularity, service type, etc.	Big data analytics helps to obtain these data, which can span across multiple layers of the network and be time-sensitive. Traditional network optimization did not use application and user data.
User data	User preference, location, mobility, online behavior, etc.	
Network data	Cell configuration, signal strength, traffic load, outage rate, inter/intracell interference, signaling, UE capability, etc.	Big data analytics is able to sense the wireless environment and the network status, thus providing globally optimized solutions. Traditional network optimization is generally confined to per radio link/user/cell optimization, or simple intercell coordination.
Link data	Physical channel information such as path loss, shadowing, channel statistics, etc.	

larity. The user data implies user behavioral profiles, such as location, mobility, and preference. Both categories can be obtained using BD analytics, for instance, through deep inspection of the user plane packets. User's location and mobility can either be obtained via GPS or network measurements. Note that BD analytics and traditional data processing may work together to obtain certain information, which may involve multiple layers in the network and have stringent time constraints. The network data includes the network configurations, such as the coordinates of the physical BS, its antenna configuration, and KPIs, such as traffic load and outage, UE capability, and various signaling interactions between UE and network. The link data concerns with the wireless links between the users and the BSs, generally obtainable via downlink and uplink measurement and reporting.

- **Potential value**
  Many mobile operators, such as CMCC, own vast ranges of WBD, including verified user information, user application usage statistics, and network data, as well as private service data. WBD is present in business, operation, and management (B-/O-/M-) domains, and on value-adding service platforms. B-domain WBD contains user's basic/personal information, user device specs, contract plans and more, which are structured data and often superior in data quality.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

O-domain WBD consists of control signaling data, network performance logs, application usage breakdown, etc. It encompasses user's online behavior, and is thus much larger in volume than B-domain data, and more difficult to process, but with greater value. M-domain WBD comprises the operator's ERP, financial reports, human resources, and alike; luckily these are normally structured data, and can provide the most valuable insights into the company's strategic financial decisions.

Operator's internal WBD has many advantages against the data from over the top (OTT) companies: 1) the data is more trustworthy since it is mandatory that mobile numbers are verified through national ID associations; 2) user profiles are multifaceted, i.e., a user's internet browsing history can be analyzed to label the individual with certain behavioral traits while a user's voice call records can reflect his/her social ties; 3) operators possess real-time information about users, such as location data; 4) data volume is unprecedented, to be specific, DPI that has captured data's daily increase is in terabytes scale.

Operators can also leverage two additional forms of data, i.e., publicly open data on the internet through web crawler systems or API on websites, and vertical industry's internal data through bilateral cooperation and data sharing.

There is also great value for terminal users through sharing information and helping with the formation of BD in order to allow the operators and service providers to provide better services and experience. Naturally, WBD is of great value to third party applications, who can leverage both internal and external data, and promote cross-association and data-mining analysis. Such cooperation can help to break free of the data island and further develop new and innovative applications and business models for the next generation. WBD analytics also provides important support for public safety. When monitoring the flow of people in real time, especially large-scale spatial location movement, it helps to find out the potential safety risk. The BS signaling data can also record the moving trajectories of users. By using user mobile sequence detection technology and key location mining algorithms along with the matching algorithm between mobile device spatial-temporal trajectories and real-time road networks, we can compute and show real-time urban traffic and provide valuable suggestions for urban road construction. Finally, wireless communication research could also benefit from WBD, in terms of physical layer transmission scheme design, channel modeling and emulation, network management and orchestration, radio resource management, cell deployment and optimization, and more.

In the past 50 years, we have seen the success of the business model of "IT+CT," which has created applications such as targeted advertisements and internet credit systems. The next 50 years will be the era of "IT+CT+DT," where BD and artificial intelligence (AI) bring new momentum to cellular industry development in terms of network optimization, capacity improvement, customized services, and better user experience, giving rise to more innovative and disruptive technologies.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

3.3.3 Artificial Intelligence in Wireless Networks

To support a range of diversified scenarios and service requirements, the 5G network is becoming softer and more agile. Unfortunately, the complexity of the network optimization problem becomes increasingly challenging if traditional methods are used. The 5G network needs to embrace new and cutting-edge technologies, such as AI, to efficiently boost both SE and EE. In one respect, ever-increasingly complicated configuration issues call for smart algorithms to replace manual changes according to prior experiences. In another respect, the network needs to intelligently adapt to the environment (e.g., traffic load, service characteristics, user behavior, etc.) to fulfill the evolving service requirements. In the following discussion, the AI concept and AI applications in wireless networking are briefly discussed.

AI is the science and engineering behind machines that behave like humans, and has long been applied to optimize communication networks [60–74]. Generally, AI encompasses multi-disciplinary techniques, such as machine learning, optimization theory, game theory, control theory, and meta-heuristics [60]. Machine learning belongs to one of the most important subfields in AI and is typically classified into three categories [75–78]: supervised learning, unsupervised learning, and reinforcement learning (RL), as shown in Fig. 3.16. Supervised learning's goal is to obtain an optimal model/function though analyzing a set of training examples, each of which consists of an input object and a desired output value. Such inferred model/function is then used to map new inputs in a seemingly intelligent way. Unsupervised learning is mainly used when the expected output is not known and the system has to learn by itself. RL works similarly to the unsupervised scenario, where a system must learn the expected output on its own, but with the help of a reward mechanism. If the decision made by the system was good, a reward is given, otherwise the system receives a penalty [79]. This reward mechanism enables the RL system to continuously update itself to maximize some notion of cumulative reward. The emerging hot topic, deep learning methods, can also fit perfectly into one of the three types [80, 83]. Figure 3.16 shows a general view of the different learning schemes [81].

Recently, ML and AI applications in cellular network domains have been heatedly discussed. In [66], a survey of ML techniques for self-organizing cellular networks was provided. In [65], typical AI algorithms were investigated to enhance cellular network performance. With the exploitation of WBD and the rapid development of learning algorithms, AI will play an unprecedented role in future networks. It helps to sense, mine, predict, and reason in the context of wireless systems, enabling fast and optimal adaptation and configuration of various network parameters/processes. AI is envisioned to be the next disruptive element for 5G and beyond to improve system performance in terms of both network capabilities and user experience. To employ AI technology for wireless communication systems, it is necessary to pay great attention to three main problems: the construction of wireless data sets, real-time requirement of network functions, and the universal applicability of algorithms. These issues need to be always kept in mind when studying any specific use cases.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

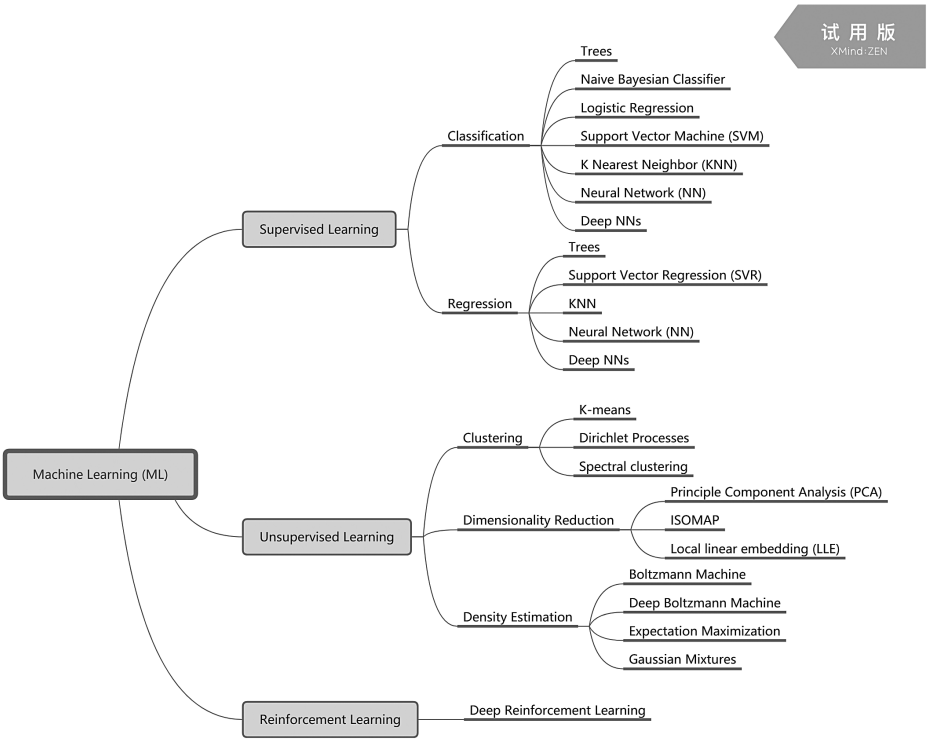

Figure 3.16 An overview of machine learning algorithms.

3.3.4 Application of WBD and AI into Mobile Network

The fundamental drive of operators' engagement in WBD research is to ensure coverage/connectivity, improve management efficiency, alleviate operational cost, maintain competitiveness, and offer superior services to customers. Traditional solutions for achieving those goals are very limited if not nonexistent, in terms of both methodology, algorithm, and implementation. With the advent of WBD and AI concepts and methods, we suddenly have a new weapon to tackle these tricky issues with promisingly high efficiency and great results. In addition, there are also AI-enabled applications where collaborative effort with third party companies can lead to new revenue for operators. A mobile operator is inherently the owner of the data, and a data service enabler, at the same time, should also strive to be a key technology holder and a data solutions provider. This ambition needs continuous research and trial and error to fulfil.

Traditional research interest in WBD has been focused internally as vital for network operators, and it falls within these four categories:

- Customer service data: Voice calls, web forums, etc. are being collected and solidified into valuable insights about the current quality of and possible improvements to our service chain.
- Sales support: By investigating customers' online behavior and forming user profiles/preferences, the success rate of targeted advertisements can be significantly improved.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

- Network optimization: Operators are using WBD analytics to better understand the relations between the network operational status and user's QoE.
- Enhanced management capability: Scientific data-mining techniques are being applied to operation and management data to improve the efficiency and effectiveness of cooperating actions. Specifically, there have been multiple success stories about using internal WBD resources on areas like wireless caching [82], network resource allocation [84], network planning, and management optimization [85], etc.

Note that the goal here has been to improve the network's energy efficiency and spectrum efficiency in the long run. For instance, the authors in [86] made an attempt on implementing a signaling-based network optimization scheme on 4G LTE networks, where such intelligent operation only happens at the mobile network management plane and is a long-term action.

A more recent development is the new approved study item in 3GPP on network data analytics (NWDA) [87], which is a network entity that is designed to provide network analytics feedback (currently only slice-specific network status) to the policy control function (PCF) at the CN. NWDA is envisioned to identify service characteristics to form a BD model, which is then used to classify incoming traffic to allow customized and improved service delivery.

At the RAN side, Fig. 3.17 illustrates some use cases ranging from PHY-layer optimization, as well as protocol and signaling simplification, to network resource orchestration, where sensing, prediction, and intelligent decision-making form the BD analytical process.

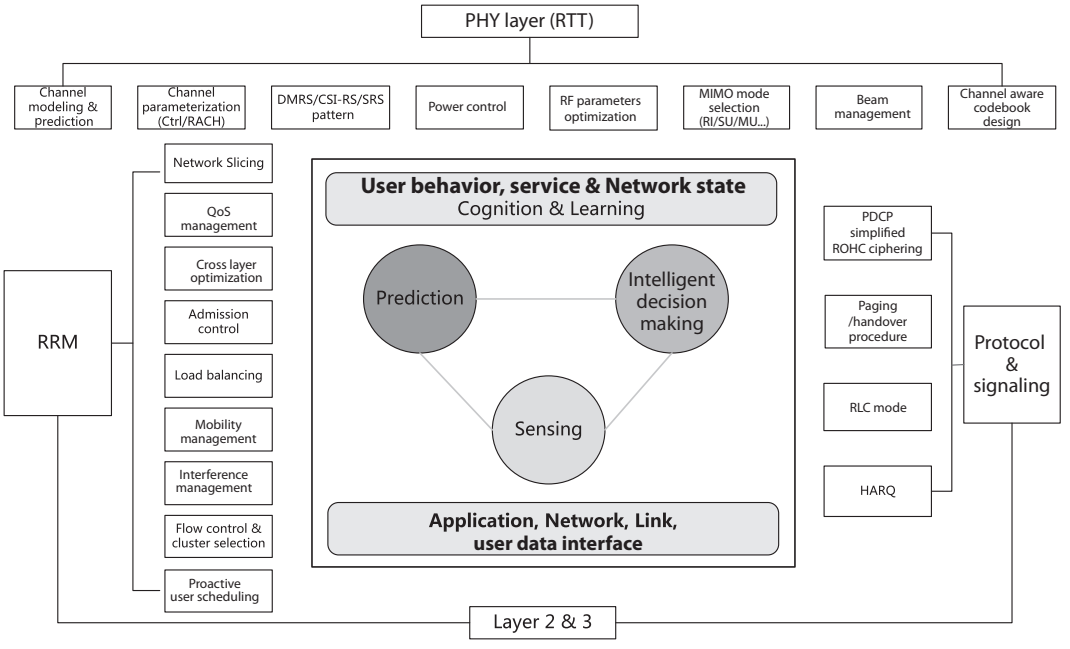

Figure 3.17 The collection of RAN optimization uses cases.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

Furthermore, various forward-thinking characterizations and use cases of WBD have been identified in [88, 89], and provided a more integrated picture of the currently fragmented research in the field, albeit mainly focused on the core network again. In [90], a BD-aware wireless network has been proposed and some state-of-the-art signal processing techniques adaptable for managing WBD traffic have been identified. Authors in [91] have further analyzed problems, such as resource management, cache server deployment, QoE modeling, and monitoring in heterogeneous networks, and proposed solutions in a wireless BD-driven framework. However, the authors have not considered how to integrate the framework into the real network. In the next section, a WBD-enabled wireless communication network is proposed, with WBD implemented in both the CN side and RAN side.

In the upcoming sections we will try to minimize the research and application gap between the CN and the RAN by introducing BD and AI concepts in many use cases and scenarios on the RAN side, along with the architecture that supports such innovation. In terms of use cases, they can be roughly categorized into four types (illustrated in Fig. 3.18).

- Intelligent NMS/MANO: Automatic network management and control, such as coverage enhancement (interference management and coverage hole discovery); QoE monitoring and optimization through CP/UP signaling and active measurements; implementation of energy-saving schemes that rely on sensing the network traffic and user profiles; cooperation and control in heterogeneous networks and among multiple RATs; cross-layer service optimization, etc.
- Intelligent MEC: MEC enables a range of new applications that require massive connectivity, huge data volume, ultra-low latency and high reliability by ways of moving the computing and radio resources closer to the users; the success of MEC applications, such as proactive caching, depends on intelligent management that is based on the understanding of the various service requirements, network status, user profiles, etc.
- Intelligent RRM: Traditional RRM includes power control and allocation, channel allocation, handover, access control, traffic load balancing, end-to-end QoS control. All can benefit greatly with intelligent setting of the control parameters and operational models based on WBD.
- Intelligent RTT: This category lies at the most fundamental level of the wireless communication system, for instance, channel modeling, spectrum mapping, signal detection, automatic MCS selection, and rank selection; the insertion of intelligence will directly improve the air interface throughput.

3.3.5 Green and Soft Network Architecture with WBD

Overall Vision of the Architecture

To enhance the system performance of wireless communication networks, we have proposed a BD-enabled network architecture in [59]. As shown in Fig. 3.19, this architecture takes different network layers into consideration, i.e., the access network, the

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

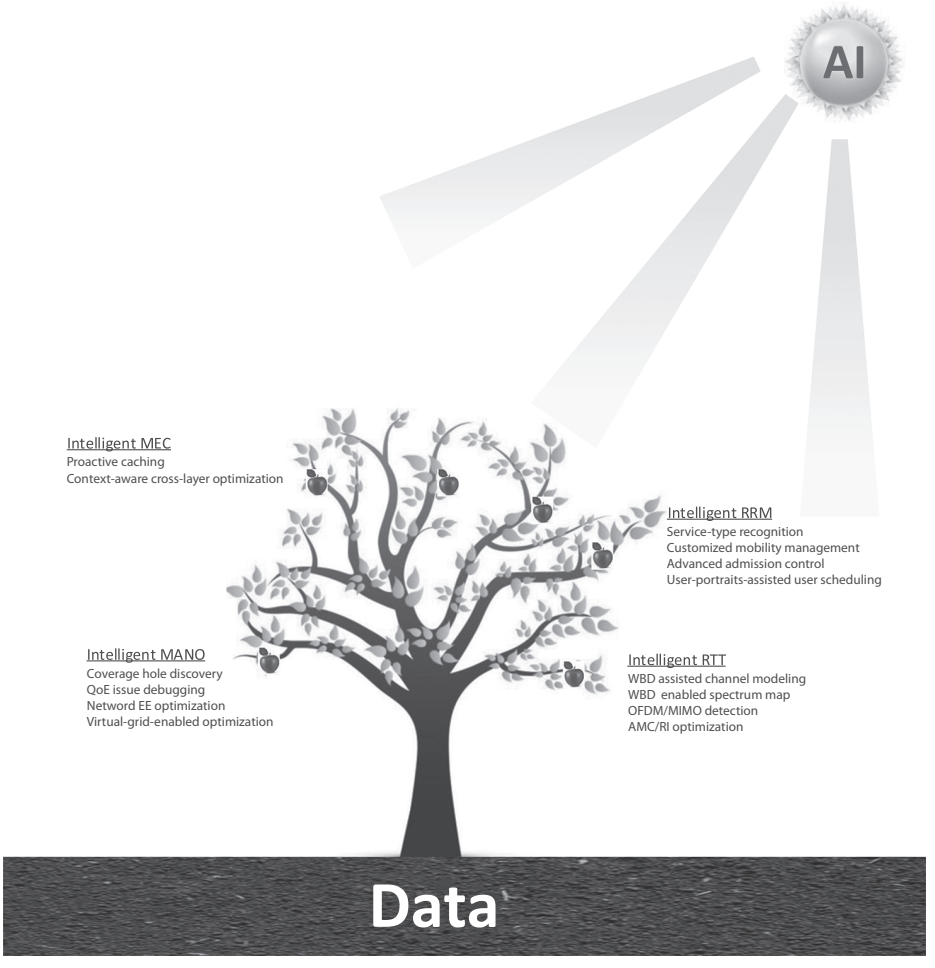

Figure 3.18 Categories of AI-driven uses cases

core network, and the IP backbone network. The distributed RAN with integrated BSs coexists in this architecture with Cloud RAN, which has a flexible functional partition between the CU and DU. Each network node, for instance, core network gateway (CN-GW), CU, DU, integrated BSs, and data-only BSs, which are equipped with low data-rate fronthaul and can only provide local data services, possesses proper storage and processing capabilities.

The BD platform is responsible for processing large volumes of data and providing useful information for resource optimization and RAN optimization. It can be deployed at either the CN or RAN side. For RAN optimization, the BD platform is recommended to be deployed at the RAN side and possibly in the CU. For CN optimization, the platform needs to be deployed at the CN. There is a possibility to define an interface between CN BD and RAN BD for information exchange. For RAN BD, the processing duty may be split between the CU and the DU. In phase 1 of 5G RAN standardization,

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

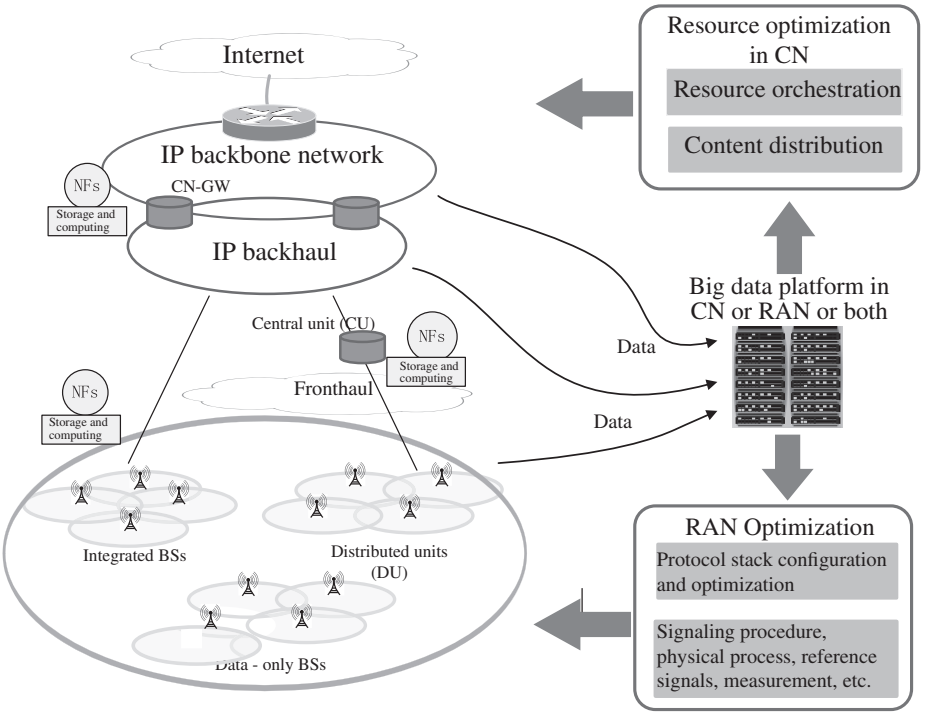

Figure 3.19 BD-enabled mobile network architecture.

the CU-DU architecture to support centralized deployment has been introduced to improve collaboration and pooling gains. Under such a centralized deployment, it is expected to deploy gNB-CU with more powerful processing and storage capabilities. Therefore, it is natural to host computation and storage intensive functionalities inside the CU, such as data collections, data storage, and data processing, including BD model training and distribution of well-trained models, etc. In the meantime, some relatively less demanding functionalities, e.g., data reporting, training-model-based decision-making, and execution, can be placed at DUs to support the real-time control. More details on the BD-empowered RAN architecture are investigated in [92, 93].

By leveraging data mining and data analytics, BD technology can help predict user mobility, traffic behavior, network load fluctuation, channel state variations, link level, and system-level interferences. This facilitates efficient resource assignment, flexible network capabilities distribution, flexible protocol stack configuration and optimization at each network node, signaling procedure, and physical layer optimization. BD changes the network design from the reactive BS-centric paradigm to the proactive user-centric paradigm.

According to the service requirements and network conditions analyzed by the BD platform, network functions (NF) implementation can be either centralized or

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

distributed at the edge. To be more specific, the new features in BD-enabled networks are listed as follows.

- **On-Demand Resource Orchestration**
  In conventional networks, a lot of resources may be wasted in light-traffic scenarios. By utilizing WBD, it is possible to predict users' future service requests, location, mobility, and the network conditions. With such useful information, a proper amount of resources can be provisioned, guaranteeing high resource utilization efficiency and reducing network cost by avoiding over-provisioning. As an example, the processing resources at the CU can be intelligently configured according to the predicted traffic fluctuation to promote more pooling gain. In addition, some access resources can be turned off if no traffic is predicted in the corresponding coverage area.

  Network slicing is a key new feature of 5G networks, defined to suit highly diversified 5G services. In order to guarantee QoS of each specific network slice, some relatively static isolation of radio resources may be adopted, which may lead to insufficient resource utilization. Different slices may have diverse characteristics, e.g., distinct peak or idle traffic patterns. There exists a possibility that the usage of radio resources for different slices may compensate for each other. For example, a slice is at its busy hour, and the other slice is at its idle hour. Such traffic patterns of different slices can be identified via BD technologies, leading to a better utilization of radio resources among slices.
- **Flexible Content Distribution**
  With the assistance of the BD technologies, the network is able to predict users' traffic patterns with more accuracy. With adequate storage and computing capabilities, the network edge nodes could pre-fetch the predicted popular contents beforehand during idle hours, instead of fetching contents upon users' request via potentially overloaded backhaul links.

  Obviously, the closer the content is to the users, the less response latency they see. However, storage and computing at higher network levels mean more pooling gain and less maintenance cost. Different traffic patterns, such as content popularity distributions, may require content deployment in different levels of network nodes. For example, when the traffic is low, or the traffic trend is of low predictability, it is more attractive to place content storage and computing at a higher level in a cost-efficient way. However, when the traffic trend can be accurately predicted, it is beneficial to improve user's experience by moving the corresponding content storage and computing functions to network nodes that are closer to users.
- **Protocol Stack Configuration and Optimization**
  With flexible function split between CU and DU, the corresponding protocol stack configuration is required. Also, there may be the scenario where DC or CA is implemented in the network. The protocol stacks at the CU, DU, integrated BS, and data-only BS need to be flexibly configured. For example, with ideal fronthaul, MAC functions can be configured at the CU, optimally allocating

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

resource between different DUs. BD analytics helps to optimize the protocol stack configuration and processing in various scenarios.

- **Signaling Procedure Optimization**
  Signaling procedure, as specified in various standards, stipulates the signaling flows between the UE, eNB, MME, SGW, PGW, PCRF, and home subscriber server (HSS), to use the terminologies of LTE as an example. With BD analytics, many signaling procedures can be simplified, which brings much-reduced operation complexity and latency and is urgently motivated in ultra-low latency applications.
- **Physical Layer Procedure Optimization**
  When the protocol stack is configured and the processing at each layer is optimized based on the service type and the scenario, the application data will be passed down, layer by layer, across the stacks. Traditional physical layer processing, such as synchronization, modulation and coding, multiple access, multiple antenna precoding, duplex mode selection, numerology configuration, reference signals, measurements and feedback, and power control, etc., can also be significantly enhanced via BD technology.

Data Driven Intelligent RAN Architecture

To leverage the WBD and AI capabilities for smart 5G, the data analytics functionalities need to be introduced into the network architecture [92]. The data analytics functionalities are mainly responsible for WBD collection, analysis, feature extraction, and model training, and provide intelligent network guidance for management and control decisions. It is quite different from the current network architecture, which mainly focuses on the communication part.

In order to meet both long-term network optimization and approximately real-time predication requirements for RRM/RTT algorithms, the architecture of the intelligence engine should be hierarchical and distributed. The overall reference architecture is illustrated in Fig. 3.20.

Multilevel BD processing and intelligent learning functions are supported. BD analytics functionalities are envisioned to be located at the operating support system (OSS)/MANO, the CN and RAN side. BD analytics will be naturally supported by the OSS/MANO for long-term network planning and network management, e.g., coverage hole discovery. The core network introduces NWDA to network data analytics (NWDA) to the 5G standard, which is used to collect massive data about each function of the core network, including user mobility data, service flow data, billing information, and network entity status data. NWDA can analyze the collected data, and improve QoS through policy control function (PCF), to enhance the user's service experience. Correspondingly, RAN domain should also define data analytics elements to optimize radio networks and coordinate with NWDA. We term this RAN data analytics element as RDA.

RDA is primarily used to support data-driven intelligent radio resource management and PHY-layer or higher-layer optimization in the RAN. The RDA should interact

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

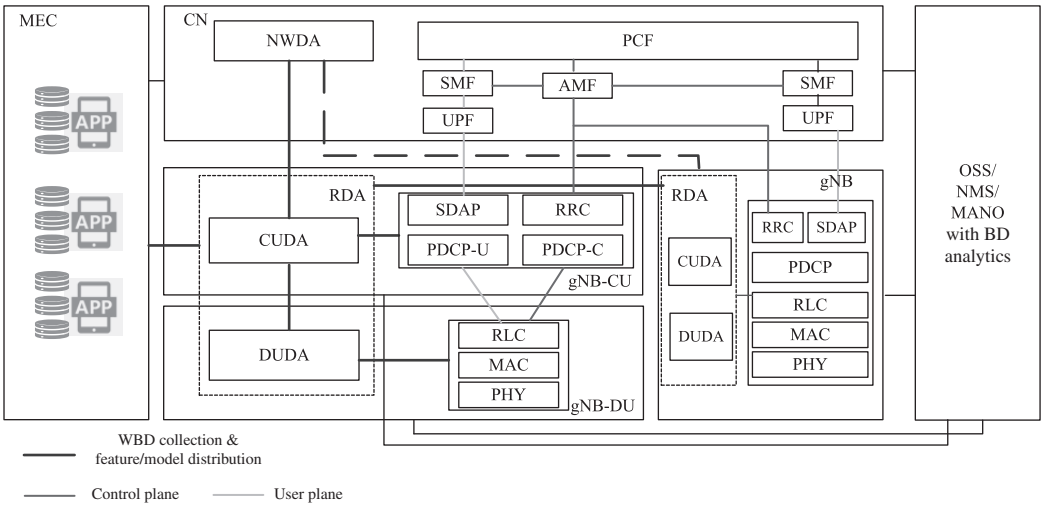

Figure 3.20 Data-driven intelligent RAN architecture

with both the control plane and user plane of RAN to realize the data collection and policy configuration. It also needs to provide data subscription services for NWDA and business and operation support system (BOSS)/OSS/MANO, and upload preprocessed subscription data to NWDA and BOSS/OSS for further BD operations and services. RDA can also subscribe to the NWDA results for the RAN-side service optimization. RDA may include central unit data analytics (CUDA) and distributed data analytics (DUDA).

CUDA is used in quasi-real-time optimization for RRC, SDAP [18], PDCP, and other protocol layers (such as multi-connection, interference management, mobility management, etc.). More specifically, it includes the data analysis, quasi-real-time predicting, decision-making model training, online model prediction, and strategy generation and configuration based on the predict results, in order to provide DUDA data features and model subscription distribution. CUDA can support both master and slave modes. The slave CUDA can request the master CUDA to perform some computational tasks, such as model training; while the master CUDA can conduct some computationally intensive model training for the slave CUDA, and offer some network-level collaborative optimization recommendations.

DUDA is adopted for real-time RAN data collection and preprocessing, prediction, parameter optimization and training tasks with low computational complexity in DU (e.g., PHY/MAC/RLC). DUDA needs to offer data features needed for training prediction/decision models after preprocessing to the CUDA, while CUDA can assist DUDA to conduct some computationally intensive model training tasks. Assisted by CUDA, the trained model can be sent to the DUDA for installation, perform real-time prediction/decision-making based on the real-time collected data, and generate

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

corresponding strategy based on the prediction results, in order to perform real-time closed-loop control for the DU's process (such as scheduling, link adaptation, etc.). Note that the architecture should also expose north interfaces to the third parties and vertical industries, and training data should be comprised of E2E network data and application data in order to provide differentiated services.

Hierarchical and distributed architecture can significantly reduce data transmission costs, since the vast amount of collected data can be analyzed and trained for local needs without being uploaded to the centralized BD analytics. Additionally, online training and prediction for real-time applications (e.g., 1ms $\sim$ 10ms) can be distributed at the DU instead of the centralized nodes with large delay backhauling, which helps to guarantee the real-time needs for the RRM/RTT use cases.

3.3.6 Big-Data-Enabled Automatic Network Management and Operation

Coverage Hole Discovery

Network KPIs (key performance indicators) can be clarified into accessibility, retainability, availablity, mobility, traffic, and radio quality. Operators use KPIs to monitor how well a network is performing. Bad radio coverage can impact network KPIs significantly. If a coverage hole could be efficiently recognized and optimized, it will improve the KPIs. It's time and man power costly to find coverage hole with drive testing. WBD platforms make coverage hole recognition easy and efficient. It's done via statistics (clustering) based on information collected in WBD platforms, such as call trace, UE measurement reporting, base station measurement results, location, and other input.

From a network point of view, it's not easy to locate UE position precisely without a positioning feature enabled. It's a fact that most networks don't enable a positioning feature. Generally, triangulation is the way used to estimate UE location. It uses base station position (preconfigured by the network operator) and direction together with UE-reported RSRP/RSRQ to do the calculation. It will work but the precision couldn't be guaranteed. Using novel machine learning algorithms, which combine elements from supervised learning-based RF fingerprinting and particle-filter-based hidden Markov model learning used for robot path-tracking [96], operators can achieve median accuracy of 20–30 m. It shows significant improvement compared to no machine learning, which is about 100 m. With this more accurate positioning info, together with call drop events, handover failure events, and RRC reestablishment events, it can quickly identify the area of weak coverage, or coverage hole. Exhibiting the result in the map, it will guide maintenance teams to perform optimization.

As shown in Fig. 3.21, red-colored dots exhibit coverage holes calculated based on UE location, UE reported RSRP, and call drop/handover failure events.

QoE Issue Debugging

Good KPIs could indicate that the network element is working at its best, but they cannot ensure meeting the end user expectations. Modern operators shift to QoE monitoring and

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

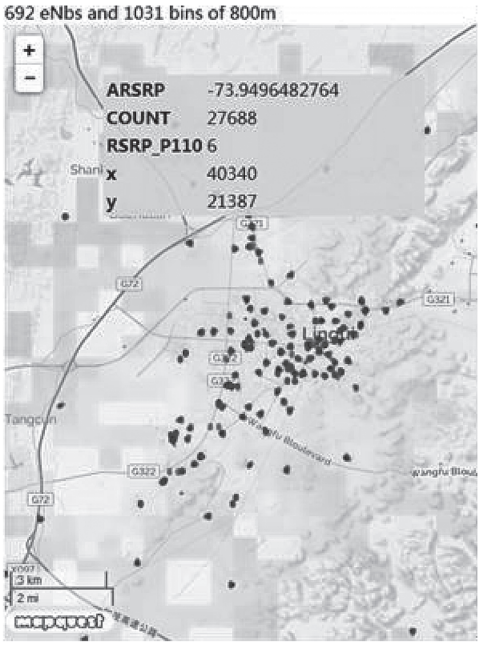

Figure 3.21 Coverage hole exhibition based on location.

optimization, which aims to improve end user satisfaction. QoE is more of a subjective experience. Factors such as personal mood could impact user perception. From the network's point of view, in order shift to QoE optimization, some KQIs (key quality indicators) are defined. KQIs could be MOS for voice, web page download time, round trip time, delay, jitter, packet loss rate, etc.

In WBD platforms, there are collections of traces from RANs and CNs, including control plane signaling, UP datas, and a variety of real-time measurement data. For a given user, it's possible to correlate information from different network elements. For example, using the unique identity of the interface in the network element, known as the time stamp, it can form an overview of a user. With all such history data and using ML algorithms, the performance problem of a given user could be automatically identified by the system, e.g., using decision trees, and the results are clustered and classified by BD analytics.

When a user is surfing the internet, if a requested web page is returned within 1s, the user feels good. If the same page were to return in more than 5s, the experience would be awful. A WBD platform can recognize all users experiencing web page download rates below a given threshold, e.g., 64 kbps for DL or 32 kbps for UL. All such users can be categorized to identify network coverage problems, device problems, terminal problems, or application server problems through decision trees, as shown in Fig. 3.22. All possible badly behaved user records are summarized and clarified; maintenance and development engineers can take this input for further checking and optimization.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

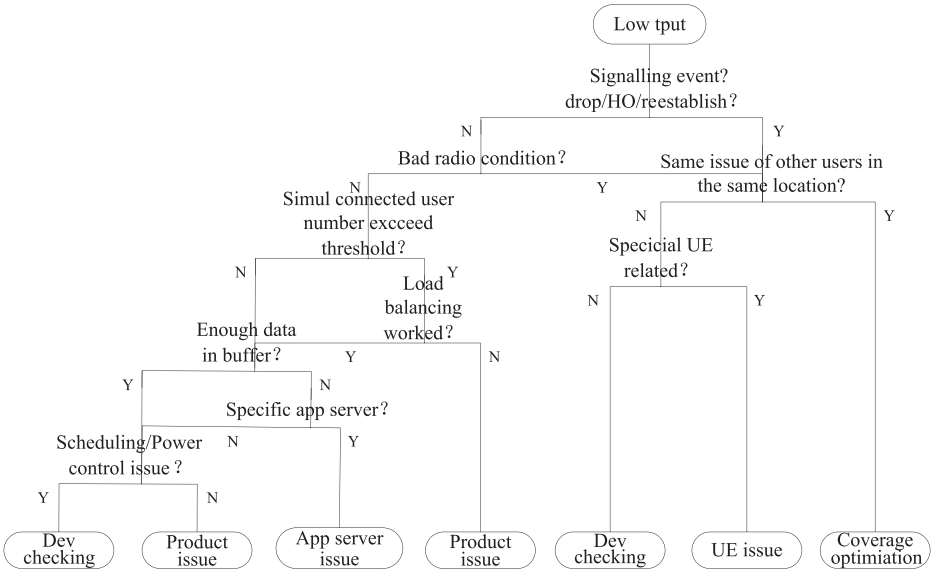

Figure 3.22 Decision tree used for low user throughput classification.

Network Energy Efficiency Optimization

Dynamic Enable/Disable Secondary Carrier

KPI data of network data volume and radio resource utilization with time of day statistics reveal that there are daytime busy hours and nighttime idle hours. It's consistent with the human's work-and-rest rhythm. There is also a significant tidal effect between office buildings and residential areas. In order to improve network coverage and user experience, operators deploy multiple carrier frequencies for hot spot. LTE advance features such as carrier aggregation can improve throughput significantly.

Power saving is one efficient means to reduce the OPEX for operators. On WBD platforms, various cycles can be achieved by the analysis of historical resource utilization data, traffic volumes. As there is also detailed application traffic information, more application behavior can be classified with periodicity and trends. With the location information, it can identify different types of coverage area for clustering. Based on the current collection of network resource usage and periodicity of change, it can provide guidance of switching ON/OFF the secondary carrier frequency. Compared to the base station method, which is based on real-time resource utilization, the number of online users to manage the extra carrier switch ON/OFF, BD platforms can more accurately predict the demand for resources and reduce the unnecessary ping-pong effect.

Figure 3.23 shows network traffic volume usage of a day for hour of a day. The time period between 23:00 and 6:00 are not shown. The secondary carrier could be dynamically switched ON/OFF based on the real network resource usage.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

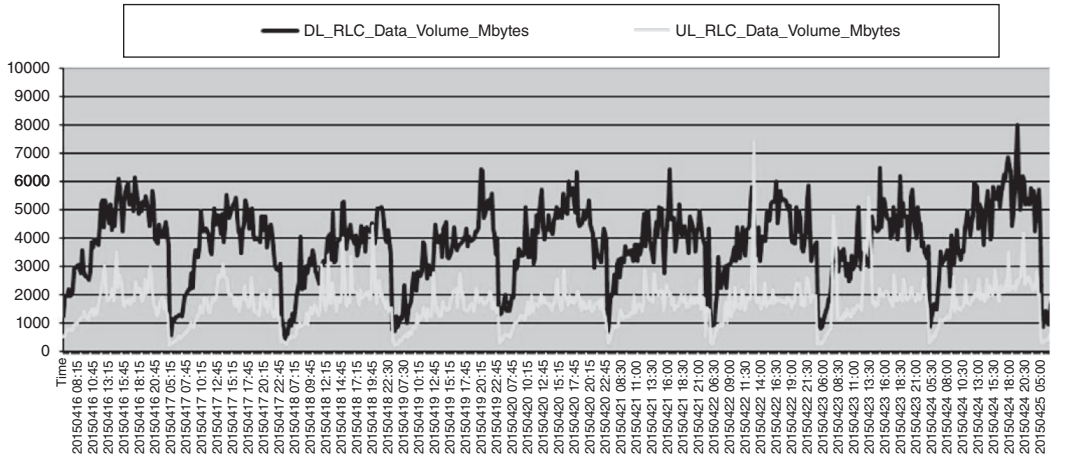

Figure 3.23 Hourly network traffic volume.

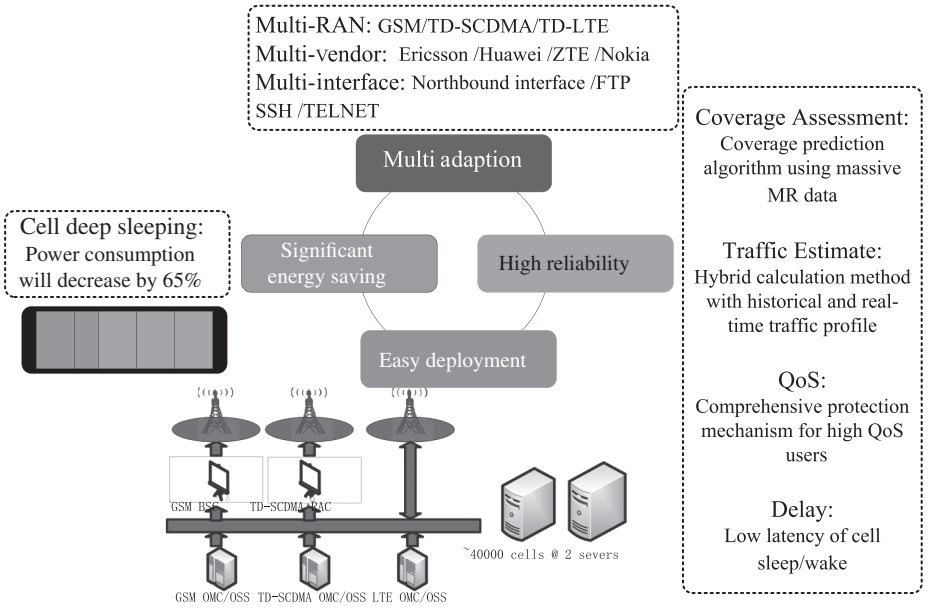

Figure 3.24 Illustrations of MCES.

Multi-RAT Cooperation Energy-Saving

To cope with the severe energy challenges caused by incessant mobile network expansion, the multi-RAT cooperation energy-saving system (MCES) was developed to effectively improve the energy efficiency of mobile networks. As shown in Fig. 3.24,

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

MCES interacts with the RAN in a real-time manner and can support multiple vendors' 2G/3G/4G RAN equipment. MCES identifies and coordinates cells with overlapping coverage to achieve network energy savings. Through BD analytics of massive MR data and traffic profiles, MCES can find the energy-saving cells and their compensating cell, and predict their traffic load trends. When the traffic load meets the criteria, for example, falling below some pre-defined threshold for some time, MCES will migrate its traffic to the compensating cell and place the energy-saving cell into a sleep state. With a real-time monitoring function, MCES can also turn on the sleeping cell before high-volume traffic arrives abruptly.

Up to now, MCES has been deployed for over 70,000 cells across ten provinces in China. The average annual electricity saving is 400,000 kwh in a 10,000-cell area.

3.3.7 Big-Data-Empowered MEC

Proactive Caching

Proactive caching predicts the contents that users may request and cache them at nodes of the wireless edge, e.g., base station (BS) and user equipment, before users send requests. The major approaches of proactive caching are caching at BSs, cache-enabled D2D communications, and pushing. When users send requests, the requested contents can be found from the cache of nearby nodes (e.g., macro BSs, pico BSs, or even user equipment) immediately and then were sent to the users, which can improve user experience [97], reduce E2E delay, and boost overall network performance efficiently [98–101].

To implement proactive caching, we need to resort to BD analytics to predict the contents to be proactively cached. In traditional wired networks, contents are stored at the content server, usually in a reactive manner. Since the content server has large storage size and covers a large number of users in a large time–space range, it is possible to predict the content popularity by BD analytics. However, in wireless networks, due to the limited size of storage and coverage area of BSs, designing caching policies based on the content popularity cannot improve caching gain efficiently. This is because the content popularity as a demand of multiple users cannot reflect the demand of each individual user, which results in lower cache-hit ratio and degrades the gain of improving network performance and user satisfaction. Therefore, to improve the gain of proactive caching, we need to predict the probability of each user requesting each file (i.e., user preference) and active level of each user. How to quickly learn user preference and active level for a large number of users with low complexity in practical environment with dynamically generated contents is an urgent task to be solved. Due to the limited number of user requests obtained at the BSS, the predicted file request probability is not accurate. Designing cache policies under these imperfect factors is also an inevitable problem.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

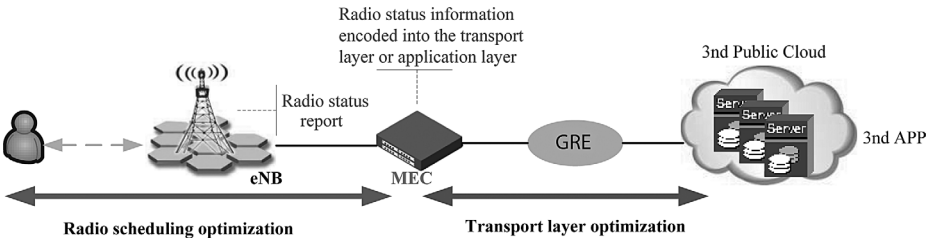

Figure 3.25 Illustration of cross-layer optimization.

RAN-Assisted Application Optimization

While the radio condition of the mobile network can fluctuate on the order of milliseconds and may occasionally result in packet losses even without network congestion, the application is adjusted on a larger timescale on the order of seconds and tends to attribute packet loss to network overloading. As a result, there is a misalignment between the RAN and the application, which may lead to noneffective usage of available radio resources and a degraded user experience. Therefore, it is highly desirable to introduce cross-layer optimization between the RAN and the application. The RAN may inform the application of the real-time radio air-interface channel status, based on which the application adjusts its transmission data rate. An illustration of the cross layer optimization is shown in Fig. 3.25.

Take TCP optimization for example, with some useful RAN information, e.g., buffer size, load of the base station (BS), the link throughput, and service type, packet error rate (PER), the TCP congestion window can be optimized and predicted to better match the radio channel variations. However, it is extremely hard to find a mathematical cross-layer model to determine the optimal TCP window with so many affecting factors. In this case, BD-assisted ML-based optimization offers an effective solution. With well constructed training data, a supervised learning-based model can be trained for the TCP window prediction. The BD-assisted learning approach allows for a good match between the TCP window and the wireless channel condition. This will significantly improve system throughput and buffer utilization.

The RAN may also identify the traffic features or traffic priorities within the application sessions via the big data analytics. Accordingly, the RAN can dynamically reconfigure its network protocols and parameters and performs prioritized scheduling strategies to guarantee traffic transmission of higher priority.

Intelligent Service Provision

The interworking between the 5G and 4G networks will be ever more tightly, especially with the introduction of NSA mode, making the network service flow more complex. How to control the flow of service and QoS, make full use of network resources, and protect the user's service experience will be some of the key factors affecting the network performance. Intelligence control of various services can be based on real-time

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

data collection (including equipment and cell size of the resource occupancy rate, user distribution, user priority, traffic distribution, and other information) together with data-mining techniques to achieve 4G and 5G network load balancing, and QoS based on user granularity.

For the NSA scenario, the intelligent control center learns the traffic distribution characteristics via data-mining algorithms, automatically allocates the traffic load on the 4G or 5G base station according to the QoS requirements of different types of services, and promotes higher throughput and resource utilization and ensures the users' experience. In addition, the intelligent control center determines through user behavioral analysis and prediction, different QoS guarantees for each user, and reserves enough transmission resources for high-priority users in advance, which encourages user loyalty. At the same time, proactive edge caching can be achieved through intelligent service control. By using MEC and coupling it with the wireless network status, the platform can learn the service profile and the best practices so as to better allocate and schedule resources according to the service priority, timing, and resource availability and further enhance the content hit rate and resource utilization efficiency.

3.3.8 Big-Data-Assisted Protocol Stack and Signaling Procedure Optimization

Protocol Stack Configuration

The traditional protocol stack for the normal integrated BS is shown in Fig. 3.26a, where the data processing is through PDCP, RLC, MAC, and physical processing, with the management of RRC. While in DC, as shown in Figs. 3.26b and 3.26c, the PDCP function resides at the master eNB (MeNB). The secondary eNB (SeNB) is only responsible for RLC, MAC, and PHY. For CA operations, joint MAC scheduling is feasible, leaving the SeNB responsible only for PHY processing, as shown in Figs. 3.26d and 3.26e. For both DC and CA, there is only one RRC at the MeNB. As shown in Fig. 3.26f, when the data-only BS is deployed with low data rate fronthual to provide local data services, TCP and IP processing can be conveniently omitted. The PDCP processing can be further simplified, e.g., IP header compression is no longer needed. With BD analytics, BSs may receive recommendations on the operation mode, e.g., CoMP, DC, or CA. The protocol stack is also configured accordingly. For example, with the information of user's geographical distribution and/or the intercell interferences, CoMP schemes may be adopted in a certain area for better performance.

In the CU/DU architecture, protocol stack partition between CU and DU can also be optimally configured via BD analytics based on service type, fronthaul capability, frequency band, user mobility, QoE, etc. Some exemplary cases are envisioned and explained as follows:

- For low latency services, fewer functions will be allocated to CU, e.g., full eNB protocol stack at DU.
- For high-frequency bands, such as mmWave, more processing is required at DU to alleviate the burden of fronthaul from extremely high data rate. One stack

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

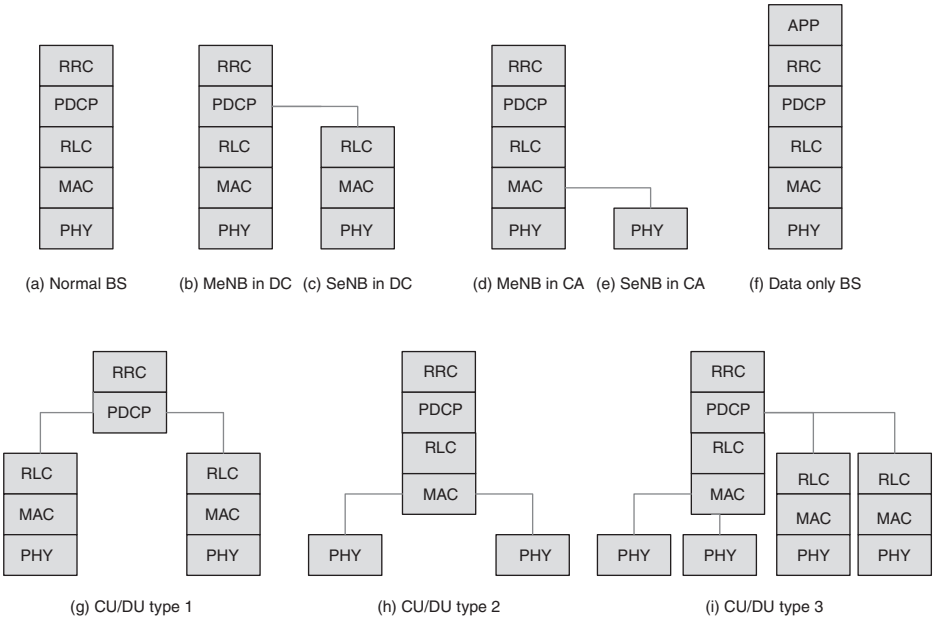

Figure 3.26 Protocol stack configuration.

configuration for cases 1 and 2 is shown in Fig. 3.26g, where RRC and PDCP reside at CU, while RLC, MAC, and PHY are handled at DUs.

- For low-frequency bands where severe intercell interference exists, more processing will be motivated at CU for efficient interference mitigation.
- With ideal fronthaul, MAC function can be configured at CU, optimally allocating resource between different DUs. While with nonideal fronthaul, MAC function at the central unit is not necessitated since cross-cell fast scheduling and resource allocation may not be well supported due to much longer fronthaul latency. In this case, only RRC and PDCP can be allocated at the CU. One stack configuration for cases 3 and 4 is shown in Fig. 3.26h, where RRC, PDCP, RLC, and MAC reside at the CU while PHY is handled at DUs. A hybrid CU/DU protocol stack is shown in Fig. 3.26i, where the CU is controlling DUs with different protocol stack configurations.
- For high-mobility users, integrated BS with a full protocol stack implementation is not preferred. Cloud RAN architecture with RRC at the CU is more suitable.

Protocol Stack Processing Optimization

- Reconfigurable Compression and Encryption in PDCP
  After the protocol stack is configured, the data processing can also be further optimized based on BD analytics. Taking a look at the robust header compression

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

(ROHC) mechanism, it is utilized to handle user plane data flow that has large packet headers; however, the incurred delay takes up 20.01% of the L2 total delay. BD analytics can be used to identify data packets or data flows with the same service types. The identified latency-insensitive IP packets can then be aggregated into a large data packet that shares one IP header, resulting in a much reduced ROHC's processing delay.

Traditional UP and CP packets all need to be encrypted when passing through the PDCP layer. The delay caused by ciphering process accounts for 59.16% of the L2 total time delay. If the service type of the packets can be analyzed and identified by BD processing, differentiated ciphering processing can be selected accordingly. If some services are not private, they do not need ciphering, thus reducing the processing delay and complexity. For example, e-commerce transactions and news browsing definitely have distinct privacy requirements. Therefore, ciphering over the air should be adaptable for different service categories to avoid unnecessary overhead. Besides, powerful BD analytics is capable of identifying potential security attacks, monitoring and eliminating potential dangers, and may effectively reduce the necessity of data ciphering.

- Optimized Transmission Mode in RLC
  Traditional RLC modes are configured by CN according to service types. Audio and video flow services basically use unacknowledged mode (UM), but acknowledged mode (AM) can be used if delay requirements are not very strict, under which reliability will be greatly improved. The UM mode transmission is more suitable for small-packet traffic as well, which usually has a small number of segments and generally does not need ARQ. BD analytics is capable of identifying traffic delay sensitivity and accurate identification of small packet traffic, bringing much-reduced delay and processing complexity.
- MAC Hybrid ARQ
  Through statistical analysis of the channel and traffic characteristics, maximum retransmission numbers can be dynamically configured by hybrid ARQ. This will bring overhead reduction and better resource utilization.

Signaling Procedure Optimization

The implementation of BD platform in wireless communication networks will naturally lead to much changes of the signaling procedure.

Taking handover as an example, we will investigate how BD helps to simplify the system operation and brings performance improvement. The basic procedure of the X2 handover [94] is illustrated in Fig. 3.27 (not including the gray color procedures with arrows and the BD center), which is very complicated. The handover is generally triggered by the eNB, based on the measurement report feedback from UEs if certain criteria of the measured channel conditions in the adjacent cells are met. The measurement process is complicated and the signaling overhead is large.

Based on BD analytics, one possible handover approach is proposed as follows (shown also in Fig. 3.27 with bold gray arrow lines). Note that the BD center in the CN is depicted in this figure, which coordinates adjacent eNBs for possibly more efficient

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

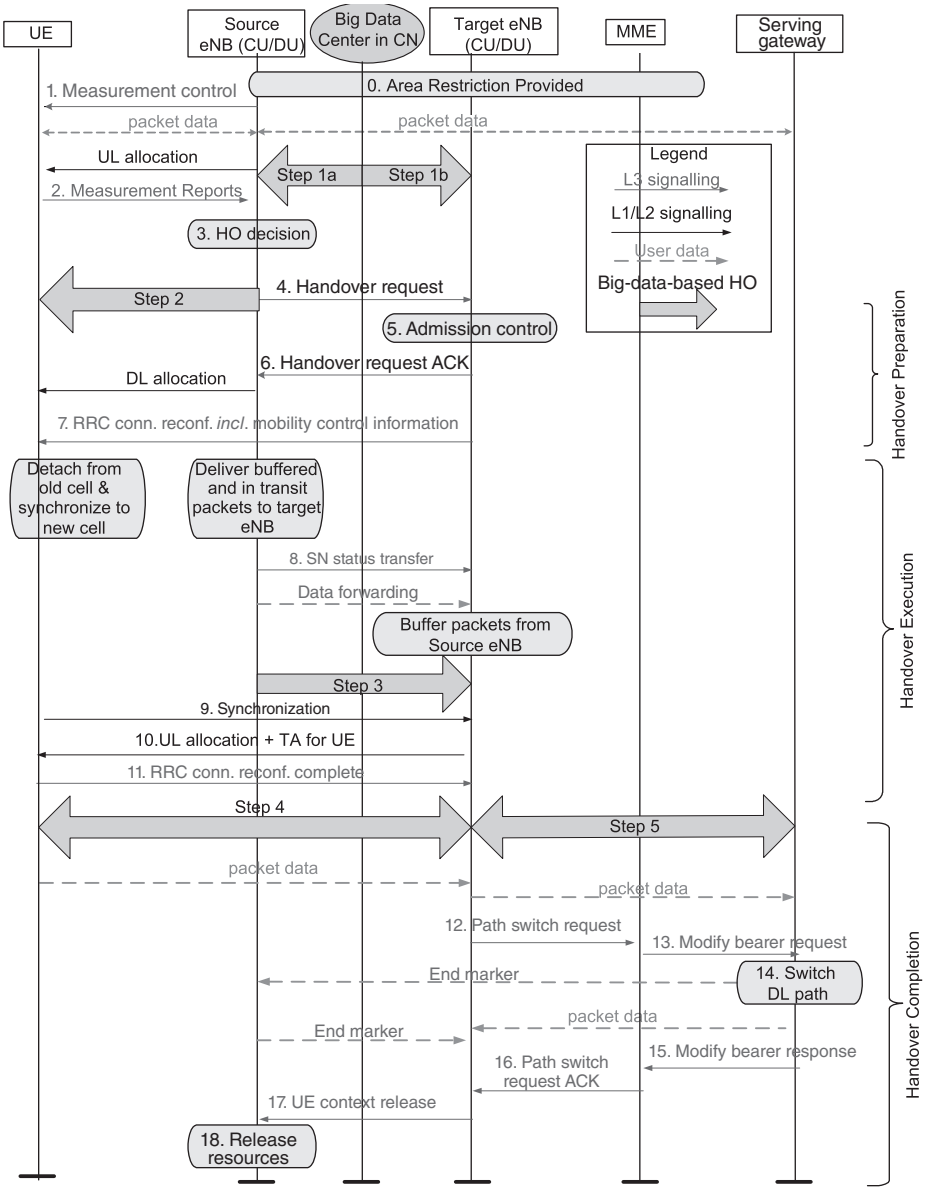

Figure 3.27 Handover procedure.

handover. For other handover cases, e.g., the CU controlled intra-DU handover, the procedure can be modified accordingly.

- Step 1a: The BD center in the CN determines when each UE should handover to which target eNB based on an accurate prediction of UE's location and

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

movement. It sends a handover command to the source eNB, with information of the target eNB.

- Step 1b: Meanwhile, the BD center sends a handover command to the target eNB, possibly with the sequence number (SN) status. Note that without BD capabilities, the handover needs multiple signaling exchanges between the source and target eNB, e.g., handover request, ACK, SN status transfer.
- Steps 2, 3, and 4 (steps 7 to 11 in the traditional handover): These can be the same as the traditional handover.
- Step 5: Path switch (steps 12 to 16 in the traditional handover) and context release can be made before steps 2, 3, and 4 finish, especially when the BD center is sure of the success of the handover.

Compared with the traditional handover procedure, BD-based handover has the following advantages: 1) Signalling overhead is reduced, e.g., handover request, acknowledgment (ACK), admission control, and possibly SN status transfer can be omitted; 2) UE's measurement and feedback efforts can be reduced; 3) Handover interruption time can be reduced due to reduced signaling procedure and to possible concurrent UE access to the target eNB and path switch process; and 4) Ping-pong handover can be effectively avoided.

This methodology can be extended to many other signaling procedures. For example, in the "attach" procedure, if it is not initial access, and the time interval between the current attach action and the previous one is small (e.g., several minutes), the identity, authentication, and security procedures can be significantly simplified or even omitted.

3.3.9 Big-Data- and AI-Enabled Radio Resource Management

Network Slicing Optimization

Next-generation (5G) wireless networks shall support various services with different characteristics. In addition to what has been supported in LTE networks (e.g., mobile broadband, VoLTE), it would support ultra-reliable and low-latency communications (URLLC) and massive machine-type communications (mMTC) services. Different services have different requirements over wireless networks, e.g., real-time video service and voice service is more sensitive to delay but has relatively loose requirements over packet loss rate; intelligent meter, like IoT devices, has strict requirement over reliability but loose requirement over bandwidth and delay. In order to meet the diverse requirements of various services, 5G introduces network slicing technique into the same physical network instruments. Network slicing is a logical concept, achieved by assigning services with the same network requirements to a slice. Different network slices are logically independent of each other and physically share the same network resources.

Per experience of traditional network planning, engineers should build traffic models and then configure resource reservation of each slice. The difficulty here is that there is

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

no such data available, and traffic model cannot reflect real resource requirement. On WBD platforms, there are collections of user traffic data (e.g., via DPI) and location info. With statistical analysis of packets, it can cluster services based on application behavior and its resource requirements, devices distribution, and periodicity. It can predict slice resource requirements by using ML and feedback to the network element for auto tuning.

Customized Mobility Management

The 5G registration area will be similar to the 4G tracking area. The registration area list is maintained in both UE and core access and mobility management function (AMF), and 5G RAN still needs to broadcast registration area code to UE. There may exist three mobility categories for UE, namely, no mobility/restricted mobility/unrestricted mobility.

Without big data, it shall be very difficult to track and categorize UEs' mobility patterns. It is beneficial to leverage BD analytics to mine the collected network information so as to precisely predict a particular UE's mobility pattern. As shown in Figure 3.28, the associated UE could track, for instance, gNB lists or cell lists per time-of-day and then feedback the earlier data analytics to the network. This allows AMF to page the UE via the reported data and therefore bring down the paging load in gNB and save corresponding processing resource in gNB.

The handover parameters have a close relationship with different network propagation environments (such as building occlusion), interference, load, and service types. If the handover parameters are set unreasonably, premature handover, late handover, or a ping-pong handover may occur, which deteriorates the handover performance and load balancing performance, and finally results in poor user experience. BD analytics can be used to collect and learn handover data in different network environments, such as different cells, slices, user types, and service types, and then to predict cell interference, as well as load and user service performances, and adaptively set handover parameter configurations.

In general, traditional methods always make handover decisions based on instantaneous measurements. BD analytics algorithms could fully exploit the historical

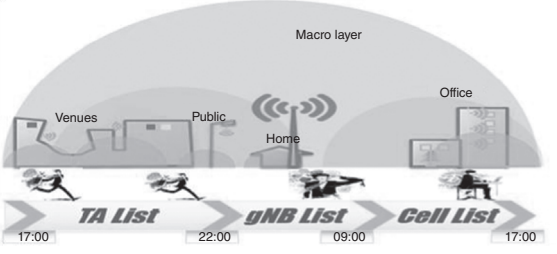

Figure 3.28 Customized mobility management.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

information and predictive information during handover procedures, which would help make intelligent decisions.

Predictive Resource Allocation

Traditional wireless systems are designed for E2E communications, which assume that the required contents, the request time, the user location, and the user trajectory are random. In these scenarios, the static resource allocation scheme is considered in wireless radio access. In other word, there is a fixed matching between the association resources (e.g., radio spectrum and BS) and RAT. As a result, the resource allocation scheme cannot be adjusted adaptively according to the dynamic tendency of spectrum and traffic load. In the network that is designed based on such a principle, the unbalanced space–time distribution of traffic causes the BSs in hot spots to overload, but there are a lot of idle resources in other BSs or in idle time.

Predictive resource allocation makes the plan for radio resources allocation in advance based on the prediction of the network environment, channel quality, and traffic demand. For example, for real-time traffic (e.g., phone calls and video conference), it is possible to ensure QoE by forecasting the traffic demands and reserving resources. For non-real-time traffic (e.g., video on demand [VoD] and file download traffic), it is possible to borrow future residual resources to improve overall network resource utilization with the individual QoE guarantee, according to user-required content, QoS requirements, and the prediction of user trajectory, channel statistics, and interference statistics. For predictive resource allocation, the related context information includes application-level information (e.g., QoE of VoD, video conference, voice, and other traffic), network-level information (e.g., congestion status, spectrum, and interference environment), user-level information (e.g., mobility trajectory and the average channel gains in corresponding locations), and device-level information (available storage capacity and remaining battery capacity of mobile terminals). So far, the existing research on predictive resource allocation considers user-level or network-level context information to allocate resources in advance, which can improve network throughput [102], reduce transmission costs, enhance user experience [102, 103], and lower network congestion [104–106].

Effective predictive resource allocation requires the prediction of user behavior, network traffic, and spectrum situation. The predicted information relies on a mass of comprehensive data from low-level layer to high-level layer, which raises a higher demand to the network function than before. On the existing network configuration, it is necessary to deploy the radio spectrum data collection module on the edge network, introduce the network traffic acquisition and recognition module at the convergence layer (e.g., the gateway), and add the matching module between the traffic information and the traffic data in the CN. Moreover, since the highly dynamic characteristic of the wireless environment causes stringent requirements on data collection, feedback, and processing cycles, it is necessary to introduce high-rate, high-reliability, and low-latency control information exchanging channels in existing network systems. In addition, the massive data generated by the dense data collection module causes serious challenges

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

on the control network, hence it is also necessary to solve the problem of all kinds of data collectors deployment.

Intelligent Network Access Control

The existing strategies of wireless device access to different-mode networks and different access points (AP) rely on a single instantaneous indicator, such as the received signal strength and the theoretical rate. The limitations of traditional access strategies in the future network have become increasingly obvious. For example, in a mmWave communication network, due to the small coverage of the network, the traditional access strategies lead to frequent switching among different APs (i.e., ping-pong effect), where redundant switching severely degrades the user's throughput. In addition, for IoT applications, diverse service-quality requirements and special device characteristics make it difficult for massive devices to access the networks quickly and easily [107]. In order to realize the intelligent access of wireless devices, instantaneous information will no longer be the only basis for decision-making; historical information and forecasted information will also be taken into account. In particular, the historical data to be utilized include network information such as received signal values, device status information, device geospatial information, throughput, and latency.

WBD contains information closely related to network behavior and the user behavior, which brings great opportunities for intelligent wireless access. With the help of BD analytics, we can extract the space–time variability and relevance of wireless service/user/network behaviors, and provide the basis for decision-making to realize intelligent and efficient wireless access [108]. There are still many problems to solve for BD-enabled wireless access, including the assessment of the service quality of different networks and the establishment of a unified assessment system. Also, the process for handling the huge feedback generated by massive devices to assist decision-making is not clear. It is also important to transfer the extracted knowledge from offline analysis of massive historical data to specific real environment for real-time online decision-making.

In order to realize WBD-enabled intelligent wireless network, the network structure needs to integrate the following modules. Firstly, an offline analysis module is required in the cloud to achieve joint mining and complex analysis of BD. High-level semantic knowledge is extracted through the complex mining of multi-modal data and will be used for the whole network. Secondly, a semi-real-time analysis module is required in the local BSs (group) to localize the off-line extracted knowledge. The module will combine the high-level semantic knowledge provided by the cloud and the local information to derive knowledge suitable for local learning. Then, an online real-time analysis module is required in edge devices to achieve real-time matching of highly dynamic networks. The module will combine the localized semantic information and real-time measured data to make quick decisions. Finally, to realize the sharing and regeneration of knowledge in the whole network, a separate knowledge flow network is required to add to the existing data transmission network, for continuously improving the intelligence of the whole network.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

Coverage and Capacity Optimization

Coverage and capacity optimization (CCO) is one of the typical operational tasks of the RAN. CCO aims to provide the required capacity in the targeted coverage areas, to minimize interference and maintain an acceptable QoS in an autonomous way. To achieve these targets, antenna power and configuration (pilot power, antenna down-tilt, antenna azimuth, or massive MIMO pattern in 5G) play a critical role, as they affect the direction of the antenna radiation pattern. Fixed RF parameters could not bring the best network performance for the ever-changing radio network environment. CCO can be used to improve the received signal strength in its own cell as well as to reduce the interference to neighboring cells by selecting appropriate RF parameters.

The network is very complicated, and it is not possible to find definite function to map between the RF parameters and the target coverage and capacity performance. The main reason is that the set of configurable RF parameters is multidimensional, and each RF parameter has a wide range of values, leading to very large number of possible options.

Reinforcement learning with the neural network is used to adjust RF parameters for CCO. The algorithm can leverage ML to analyze and learn what the proper action is for each current network state. A neural network is used to build the mapping between network states, RF parameters, and the network performance. Compared to the traditional Q-table method, it has better generalization ability and can respond to changes in the wireless network in a good way. The algorithm also punishes KPI violations. This policy guarantees the network KPI remaining stable during the process of CCO.

RF optimization based on CCO can be used in many scenarios, such as carrier aggregation (CA), energy saving, and SFN. A CA scenario is present in Fig. 3.29 by way of example. Nowadays, more and more terminals begin to support multiband CA characteristics, which makes CA terminals fully use multiband network resources, improving terminal throughput and other performance. Because the different frequency bands are not related to coordination or antenna pattern, the CA terminal of the CA wireless environment is different. For example, some terminals are in the poor CA area (low-quality CA area) in multiple frequency bands and RSRP, although giving multiple frequency bands resources, the overall throughput of current CA users is not very large.

CCO-based RF optimization is a scenario-oriented CA feature-enabling algorithm, which improves the capacity upper bound of CA characteristics in the multiband region, and forms a good complement to both the RRM and CA characteristics.

The future of CCO-based intelligent RF will involve scenario-oriented RRM characteristics to constantly improve their theoretical upper bound.

Virtual-Grid-Enabled Network Performance Optimization

Compared to traditional geographic grid, Virtual Grid does not need to divide the grid according to actual locations. Instead, the system measurements (e.g., RSRP) of multiple cells are used to define the grid. Historical KPI statistics were stored in Virtual Grid, and used for grid-level radio performance optimization. The concept of virtual grid is illustrated in Fig. 3.30, where the grid is defined by the RSRP measurements from three adjacent cells.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

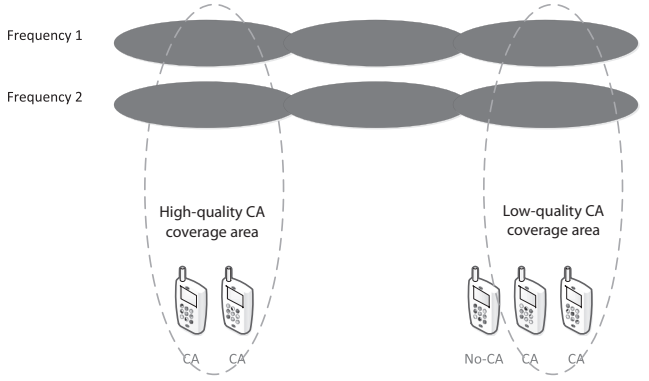

Figure 3.29 AI RF enabling CA roof.

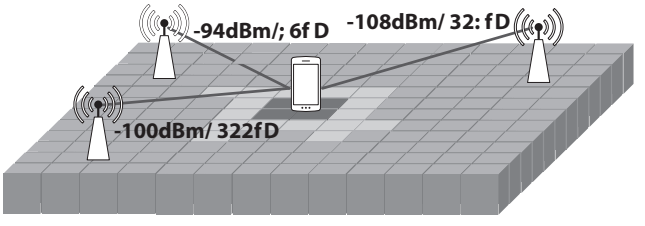

Figure 3.30 Virtual raster example.

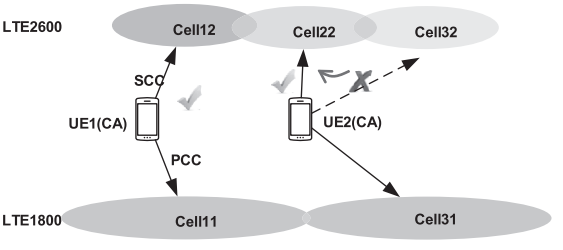

Figure 3.31 SCC selections in CA.

In the inter-frequency, inter-systems and multifrequency network scenarios, changing the granularity of the current cell-level algorithm to the grid level can improve the performance of many features such as CA, MLB, inter-frequency and inter-systems handover CSFB, and SRVCC. For example, in a CA scenario, it usually configures SCC in a blind way to reduce the GAP measure of inter-frequency, however, the carrier or PCI selected may not be the best, and sometimes may even fail to configure. As shown in Fig. 3.31, UE1 and UE2 will select Cell12 as SCC when using blind configure, but UE2 may fail. UE1 and UE2 will select Cell12 and Cell22 as SCC respectively when using the Virtual Grid method.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

Figure 3.32 Wireless network user portraits.

For BD-driven methods, the system measurements of multiple cells act as wireless fingerprints, which are used to associate the statistics and measurements stored in the grid. After that, the model can be built and then be used to predict desired statistics and measurements given the inter-frequency measurements. By doing so, suitable policy and action can be selected to improve the performance of many radio features.

User-Portraits-Based User Scheduling

User portraits refer to the classifiaction of users by given features based on real data. Through user classification by wireless network features and data, the wireless network can be targeted on-demand to provide the appropriate services to improve the user experience and network utilization.

There are many dimensions that describe user features in the wireless network. For example, as illustrated in Fig. 3.32, RSRP, RSRQ, and CQI could describe the characteristics of the wireless environment; terminal type, chip type, and transmission capacity could describe the characteristics of the user type; service type and buffer length could describe features of user demands; some features like location, movement speed, and trajectory could be used for further analysis.

For different wireless applications, we can choose different multidimensional feature combinations to classify users. For example, you can sort the user transmission priorities by the service type and buffer length; assign the appropriate cell access and presence policy to the user through the service type and the movement speed; user location, moving speed, motion trajectory, or business type to find appropriate user pairing.

3.3.10 Big-Data-Assisted High-Efficiency Physical Layer Operation

Channel State Information (CSI) Acquisition and Feedback

The availability of CSI at BS and UE is crucial for wireless communication systems. To this end, reference signals are widely adopted in various wireless standards for estimation of fast-changing wireless channels. For time-varying channels, the wide-sense stationary uncorrelated scatter (WSSUS) model is widely used in theoretical analysis. Typical channel responses in a multiple path environment in time domain (i.e., $h(t, \tau)$,

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

where, t and τ denote time and delay spread, respectively) and in frequency domain (i.e., $h(t, f)$, where f denotes frequency) are all very dynamic, especially in high mobility scenarios.

The density of reference signals in frequency division duplex (FDD) systems should be high enough to capture the channel's characteristics in both time and frequency domains. Generally, this density is selected to accommodate high mobility and large delay spread, especially for the cell's common reference signals. In cases where most users within the cell are moving slowly, a high density of the reference signals leads to unnecessary overhead, which is over-provisioning. The situation is more problematic with a large number of antennas at the BS, since the overhead of reference signals scales with the number of antenna ports.

One efficient way to minimize the overhead of the reference signals is to resort to WBD analytics. With the prediction of users' mobility and wireless channels' statistics, the BS is able to schedule common reference signals accordingly. If there are no high-mobility users, very sparse reference signals may suffice. If high-mobility and low-mobility users coexist, the BS can schedule dense reference signals in some frequency bands and sparse reference signals in other bands, thus effectively minimizing the overall overhead.

In time division duplex (TDD) systems, the CSI can be obtained by exploiting the channel reciprocity using uplink pilots. However, due to limited training sequences, pilot contamination may dramatically degrade downlink performance, since the uplink channels are contaminated by intercell interference from the adjacent cells. In order to reduce pilot contamination, with the help of BD, joint user and pilot scheduling schemes can coordinate the non-orthogonal pilot sequence among users of weak mutual interference, and arrange orthogonal pilot sequences to users with strong mutual interference.

Transform-Domain Signal Processing for Semi-Static Scheduling

Dynamic channel variations in the time and frequency domain necessitate dense reference signals, fast feedback, adaptive modulation and coding, and fast scheduling, thus posing tough challenges for wireless communication system design. Fortunately, via Fourier transformation with respect to variable t, the dynamic $h(t, \tau)$ can be transformed to a stationary $h(v, \tau)$ in delay and Doppler domains, where v is Doppler frequency. In contrast to the time and frequency domain channel responses, the delay and Doppler domain channels are more stable, depending merely on the multipath channel structure (angular distribution and the power delay profile) and mobility. Therefore, it is almost static if the channel structure and mobility does not change.

The direct impact of the transform-domain signal processing is the alleviated difficulty in tracking the time-varying fading. This is particularly useful in high-speed train communications. The significantly increased coherence time of the effective channel in transform domains brings abundant opportunities for the simplification of wireless systems in both the standardization and the implementation. For example, reference signals can be designed with very low overhead and the channel feedback need not to be fast. Channel coding schemes can also be simplified. The well-studied AWGN codes

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

may perform sufficiently well over the effective channel, thus alleviating the burden of traditional adaptive modulation and coding. Moreover, it enables FDD massive MIMO in moving applications due to easy CSI estimation. Most importantly, the slow variation of the effective channels in transform domains can significantly facilitate BD analytics since analyzing the statistical channels may be enough for satisfactory PHY and MAC operations.

Flexible Frame Structure Configuration

Since there may be many use cases emerging in 5G and beyond, it is very important for operators to deploy one network to support all deployment scenarios and use cases. Toward this end, it is critical to adopt one unified and flexible air interface framework to meet diverse requirements of the key 5G scenarios, e.g., eMBB, URLLC, and mMTC. The unified framework of SDAI [95] may include flexible frame structure, waveform, duplex mode, multiple access, MIMO, coding and modulation, and corresponding layer 2 and layer 3 signaling. For efficient operation of SDAI, the key is a flexible frame structure, e.g., the numerology can be dynamically configured; the time resource within each subframe can be flexibly partitioned between downlink and uplink; the duration of the transmit time interval is adaptive; and the period of uplink feedback of ACK/NACK is configurable. The practical implementation of flexible frame structure at each BS is very difficult. For example, flexible downlink and uplink transmission may bring severe intercell and intracell interference. However, mitigating interference, especially the cross-link interference, can be very challenging. Another example is that the frequent uplink feedback for the latency-sensitive service may cause severe interference to the downlink of the latency-nonsensitive service. Thanks to BD technologies, a lot of useful information can be utilized to optimize the frame structure, e.g., UE's service types, QoE, location, traffic volume, mobility, channel information, and inter-user interference.

Power Control

In mobile communication systems, the power control mechanism is mainly used in uplink to adjust UEs' transmission power for the purposes of compensating time-varying path loss and reducing mutual interference. The power control procedure of LTE uplink data transmission generally consists of two parts, basic open-loop operating point and dynamic power offset.

The basic open-loop operating point is determined by UE bandwidth M, preconfigured cell nominal power P_{o_PUSCH}, estimated downlink path loss PL and path loss factor a. The BD technique could potentially aid the value setting of certain parameters. For instance, the path loss factor is chosen to balance the edge UEs' performance.

In the downlink, eNB marginalizes the functionality of power control, and only adopts power allocation between downlink traffic data and control signaling. By utilizing network load data, we can optimally allocate the power ratio to maximize the network energy efficiency under the constraint of various QoS requirement of UEs. Furthermore, benefiting from the accurate estimation of channel state information and characteristic of deployment scenarios, a more flexible downlink power control scheme may be also considered in future mobile network design, e.g., interference map-based power control.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

Machine-Learning-Based MIMO Transmission

In those large-scale MIMO-equipped future wireless communication scenarios, the computational complexity in determining optimal transmission using matrix calculations becomes unfeasible. In particular, when combined with multi-carrier technologies such as OFDM, the optimal transmission mode cannot be obtained in closed form. Applying ML on massive MIMO link adaptation can deeply exploit the inherent connection among channel characteristics, transmission mode, and error performance. Based on data-driven ML methods, the mapping relation can be effectively learned by observing the channel data and the corresponding error performance so as to select the optimal transmission mode for each channel realization. It is of vital significance to apply the combination of deep learning and classification algorithms in ML on feature extraction of the channel state information so that it can improve the adaptive learning ability and universality of the MIMO link, further reducing the computational complexity.

3.3.11 Big Data Platform Capabilities/Environment

Platform Requirements and Definitions

To support fast-evolving, data-driven wireless services, the following five capabilities are required for the WBD platforms: 1) BD clusters management capability; 2) BD analytics capability; 3) wireless networks, wireless transmissions, and wireless applications support capability; 4) self-optimization, iterative update, and continuous integration capability; and 5) the AI/ML framework support capability.

The key features of WBD platforms are summarized as follows:

1. **Simple and unified management interface:** The management interface is expected to effectively support the WBD platform construction, operation, and maintenance services. It needs to support forming a distributed WBD platform via the management interface. In addition, different BD software and an integrated BD capacity platform can be built according to the different hardware resources of components. For the user interface, powerful engineering management capabilities and unified operating interface are desired.
2. **Efficient off-line data analysis:** To efficiently deal with the massive amount of structured, semi-structured, and unstructured data, the WBD platform needs to be able to do unified processing for different types of different systems. In addition, specific efforts are also needed for fast and stable processing of WBD.
3. **Real-time online data analysis:** For some wireless applications, the platform needs to provide real-time analysis capabilities to make timely decisions. The timescale for the real-time decision can be by the hour, minute, second, or even millisecond.
4. **Support off-line algorithm:** Data analysis usually provides an elementary statistical analysis. For a deeper analysis, a ML algorithm may be needed to build a dedicated decision model. The WBD platform needs to provide the off-line algorithm for the original off-line data analysis, e.g., model training. The off-line

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

algorithm mainly refers to non-incremental algorithms. Taking the wireless network optimization use cases for example, instead of building complex mathematical models through manual experience, the WBD platform helps to build a data-driven model, which is expected to obtain a more accurate and practical mapping, and automatically provide suggestions.

5. **Support online algorithm:** Online data analysis provides real-time feedback for fast and timely response and decision-making. The online algorithm mainly focuses on the incremental algorithm or the reinforcement algorithm that can make the previously constructed model evolve in real time. It helps to reduce the human intervention and avoid waste.

Framework of Wireless Big Data Platforms

A WBD platform can be built based on the cloud or physical servers. The framework of the WBD platform is shown in Fig. 3.33. The platform is divided into three layers, namely the data acquisition and preprocessing layer, the BD computing layer, and the application/algorithm layer. It uses the corresponding interface between the different components for isolation.

For the data acquisition and preprocessing layer, we mainly consider the collection of the commercial network data, laboratory simulation data, and third-party data. Data transmission includes both off-line transmission and online transmission according to the application requirements. As the data form is very diverse, a unified representation

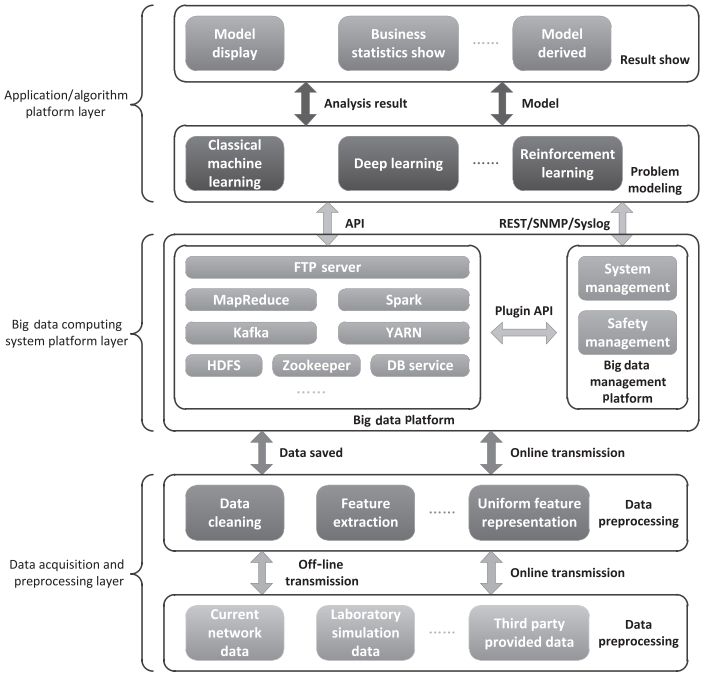

Figure 3.33 Framework of a WBD platform.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

is desired. Moreover, we need to preprocess the data by cleaning and feature extraction before using it for specific applications. With regard to specific use cases, the data can be stored in a BD database or directly input to the algorithms.

The BD computing system platform layer is composed of the BD analytics part and management part. The management part is mainly responsible for BD platform operation and maintenance, platform capacity expansion, etc. As for the BD analytic part, we need to take account of the wireless services' requirements on the platform, and the capability of the different physical servers. For example, a machine that computes superior performance can deploy Spark services to support online and off-line computing. While machines with sufficient storage space and better performance exist, Spark and HDFS services can be deployed to support off-line processing of massive amounts of data. As communication between servers often limits the real-time capabilities of cluster computing in BD clusters, it is recommended to use fiber, optical switching, and other more rapid communication media to speed up the communication between the nodes.

For the application layer/algorithm platform, the following three aspects should be considered. The first is the project management capabilities. It is required to be friendly to users with different backgrounds. The second is the algorithm. Many outstanding ML libraries can be integrated. The ML algorithm can be divided into classical ML algorithms, incremental learning algorithms, deep learning, and reinforcement learning. For classical ML algorithms, the needs of stand-alone and distributed clusters must be taken into account, blending in such tools as scikit-learn, H2O.ai, Spark MLlib, and so on. Incremental learning can refer to the StreamDM system developed by Huawei Noah Labs. For deep learning algorithms, the platform can be integrated with Tensorflow, Caffe, Theano, and other deep learning systems. For reinforcement learning algorithms, Caffe2, DeeR, Tensorflow, and other existing system frameworks can be considered. The third is the efficient communication to support real-time online analysis and algorithm construction for wireless service.

Data Representation in Wireless Big Data Platforms

The data processing performance of the platform heavily relies on the data representation. It not only affects the storage of data, but also affects the efficiency of data processing. Here, we provide some guidance for the data representation of different data structures.

In general, data can be divided into structured data, semi-structured data, and unstructured data. A two-dimensional table is usually used to express the structured data. Call history record, words, KPI data from the commercial network, or laboratory simulation can be regarded as structured data. It is not convenient to present unstructured data as a two-dimensional logic table. It generally includes all formats of office documents, text, pictures, images, audio/video, and so on. Compared with structured data, unstructured data usually has unknown data semantic information and requires a large amount of space for single-record storage. Therefore, it is difficult to store and retrieve the unstructured data. Semi-structured data exists between fully structured and completely unstructured data, such as HTML documents. It is generally self-describing; the structure and content are mixed together.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

For unstructured data, it is recommended to use a table containing the metadata of three fields to represent: number, description (varchar (1024)), content index (blob). Metadata uses structured data representation to manage the source data. Through the content description, we can know the meaning of the data. Through the content index, we can obtain the corresponding raw data. For the storage and representation of semi-structured data, two widely used methods are suggested:

1. Decompose the semi-structured data into structured data by extracting the the required fields. If the data does not contain certain attributes, the vacancy or default value can be used. The rest of the data can be added to the memo field as the note information. This method is convenient for querying. But it lacks flexibility of expansion, which is unable to handle the extra information beyond the preliminary design.
2. Use XML format to organize the unstructured data and save to the CLOB field. Information of different categories can be stored in different nodes of XML. The advantages of this approach are: it is easy to expand the information by simply changing the corresponding DTD or XSD. Disadvantages are the relatively low query efficiency.

3.3.12 Enhanced System Performance with WBD

In the previous three sections, we have discussed how WBD can impact wireless communication network design, from different perspectives, such as network architecture, protocol stack, signaling procedure, and physical layer operations. A BD-enabled network architecture has been proposed, along with design considerations on protocol stack configuration and simplified signaling procedures such as handover, simplified stack processing in PDCP, RLC, and MAC layer, low overhead reference signals, and flexible frame structure. The potential impact of transform domain signal processing on system design is also discussed, which facilitates the application of WBD. WBD, which is generated in mobile networks and is seemingly a burden to the network, nevertheless can be eventually transformed to a blessing, enabling much simplified network operations and standardization. In this section, the enhanced system performance with WBD is investigated with some preliminary simulation results. We will present three typical use cases in the wireless domain [92], i.e., mobility management, TCP window adjustment, and beam sweeping procedure, where BD and AI algorithms are used to optimize simulated system performance in a heuristic manner.

Big-Data-Enabled Efficient Mobility Management

With BD, user behavior such as trajectory pattern can be accurately predicted, which helps to improve the efficiency of the mobility management in cellular networks. In LTE, mobility management includes the paging procedure in the idle state and the handover procedure in the connected state. In the following section, the efficient mobility management scheme based on the BD analytics is discussed for handover procedures.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

When UE is at the connected state, mobility management is realized by the handover procedure. Handover optimization has two sometimes conflicting objectives: minimizing the unnecessary handovers and minimizing the likelihood of dropped calls. The state-of-the-art optimization method is mainly based on two tunable parameters: time to trigger (TTT) and handover hysterisis value (Hys) [109], which is also known as the mobility robustness optimization studied in the self-organized network (SON). The handover parameters optimization is a complicated task since the coverage areas of BSs are usually irregular and the signal quality and noise vary rapidly due to the variation of shadowing and multipath propagation in realistic wireless environments. When TTT and Hys are set too small, unnecessary handover may occur and induce ping-pong effect. When TTT and Hys are set too large, it will increase the probability of radio link failure.

Handover Parameters Optimization

The traditional handover is manually carried out by network operators and based on prior experience. Recently, BD analytics has been introduced to facilitate the automation of the mobility robustness optimization in SON. Via BD analytics on the historical handover KPI, the optimized cell-specific TTT and Hys can be found via either statistical analysis or advanced reinforcement learning algorithms. To further improve handover performance, the parameters can be further enhanced to be a user-location-specific rather than the rough-cell-specific basis. As shown in [110], the number of dropped calls and the number of unnecessary handovers can be significantly reduced via location-specific handover parameter settings. Through an unsupervised neural network, the simulated scenario of the specific indoor environment is efficiently learned, and the handover parameters are finely tuned.

Another promising handover optimization direction is to rely on BD analytics to automatically find service-specific handover parameter configurations. Now 4G RAN is configured with two fixed handover thresholds for VoLTE and data respectively. However, operators have to perform field tests to get the voice handover threshold, which results in high OPEX as well as long time to market. Besides VoLTE, more new services such as V2X will be supported and likely to have their own handover threshold preference. Moreover, service coverage could be very different due to service usage pattern, e.g., temporal factors. During the daytime more traffic (e.g., web browsing) shows up in work places and transportation hubs, and during the night more traffic (e.g., video) appears in residential areas. To perfectly match the traffic and service variations, BD analytics is in need. By leverage BD analytics on the coverage and service performance, RAN can derive appropriate service handover threshold accordingly. Thus, the occurrence of ping-pong handovers and handover failures will drop, leading to better user experience and reduced signaling overhead.

Intelligent Handover Decision Policy

In some specific scenarios, the conflicting demands of minimizing unnecessary handovers and the likelihood of dropped calls cannot be satisfied using the traditional handover method, i.e., adjusting TTTs and Hys of the serving cell and the neighbor cell.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

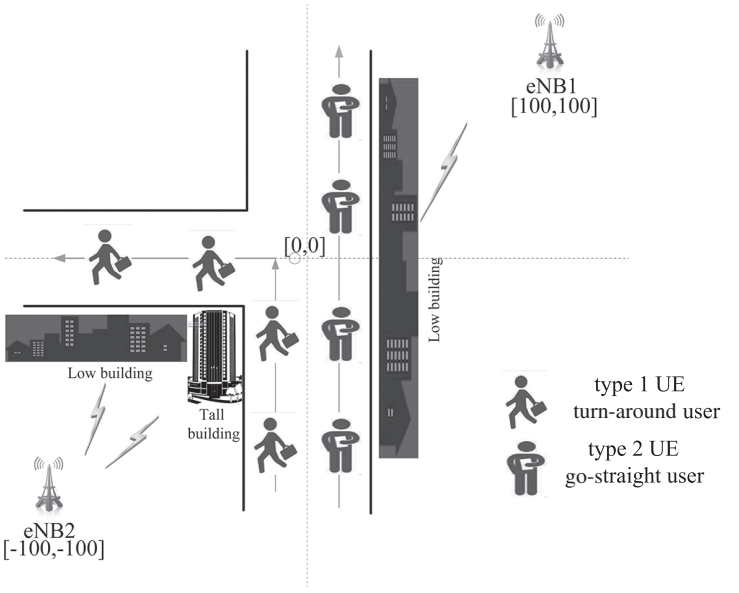

Figure 3.34 Example of the handover scenario.

A handover scenario is illustrated in Fig. 3.34. Two eNBs are respectively deployed at [−100,−100]m and [100,100]m at a crossroad. One tall tower stands at the turn corner, and two low buildings are along the street. In this example, we deployed two types of users, type 1 UE (UE1, turn-around) and type UE 2 (UE2, go-straight). At the initial stage, UE1 and UE2 were both connected to eNB2 due to higher reference signal receiving power (RSRP). For UE1, his eNB2 RSPR experienced a decrease first, followed by an increase after making the right turn. To avoid ping-pong handovers, it was better to set a large TTT. For UE2, his eNB2 RSRP decreased along the route, and the best policy was to handover from eNB2 to eNB1. If TTT is set to a large value, the probability of dropped calls will increase for UE2. In this case, the appropriate TTT settings for UE1 and UE2 are in conflict, which the state-of-the-art approach cannot efficiently resolve. For this case, more intelligent handover decision policy is needed.

The data-driven learning-based algorithm emerges as a potentially fantastic solution. Based on the historical RSRP, reference signal receiving quality (RSRQ), the handover records, and the corresponding handover performance, e.g., throughput during the handover, powerful machine learning techniques are able to learn the realistic propagation environment. An intelligent handover decision policy is proposed by solving a classification problem via supervised machine learning. To do so, a classifier can be built based on the user's RSRP/RSRQ sequences along the route. The RSRP sequence serves as the input vector, and the output label is the best handover cell index based on the judgment of the handover performance. As for the simple example here, labels can be simplified as 1 or 0 for handover from eNB2 to eNB1 or not. Support vector machines (SVM) are used as the ML technique to perform the classification. SVM seek for the optimal hyperplane,

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

which optimally separates the data into two classes in the feature space. The proposed data-driven learning-based handover scheme includes the following steps: 1) training dataset generation; 2) build a training model and model testing; 3) model execution in real-time handover procedure.

1. Training dataset generation
   - Input samples:
     The training input samples are also known as features. The RSRP, RSRQ sequence extracted from the measurement report (MR) can be selected as the training input samples to capture the feature of user's trajectory.
   - Labeling:
     The labels are generated based on the evaluation of the handover performance. In general, the labels are set as the cell index that gives the best handover performance. For instance, if the ping-pong handover is detected for certain input vector, the label will be set as the current cell index to indicate that it is better to stay on the current cell rather than perform handover.
2. Build a learning model
   Using the labeled training data, the SVM classifiers can be obtained.
3. Real-time handover decision making:
   With the trained classifiers and the real-time input of the MR data, intelligent handover decisions can be made.

   In Table 3.2, we show the count of service interruption and the count of the ping-pong handover for 100 simulation instances. In this simulation, the ping-pong handover is defined as a handover from cell B to cell A, then handover back to cell B if the time-of-stay-connected in cell A is less than 1s. The service interruption happens when the user's signal-to-interference-and-noise ratio (SINR) of served cell is less than 3dB during a period of 100ms. In that case, the effective transmission SINR will be zero, due to data transmission failure. For each instance, we generate one user, and that user may go straight or turn around with the same probability. The result shows that the proposed handover scheme can efficiently reduce ping-pong handovers and service interruptions. It indicates that data-driven learning-based optimization can be used to improve users' mobility performance. It will be especially useful in more challenging propagation scenarios. Note that the proposed algorithm can also be easily extended to wider areas and more complex deployment scenarios (e.g., heterogeneous cells).

To further enhance the learning-based handover scheme, the probability of the user's direction of movement can be designed as the additional input of the handover classifier trained by the supervised ML algorithm. To support the acquisition of such probability, the CN BD platform may extract the mobility profile of the user by analyzing his MR data over the duration of days/weeks/months and obtain the probability at each cell with respect to time, then push the user mobility profile to the corresponding CU when that user enters the cell hosted by that CU.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

Table 3.2 Counts of service interruption and ping-pong handover for 100 instances

TTT setting (ms)		*200*	*400*	*600*
State-of-the-art method	*Count of service interruption*	7	26	51
	Count of ping-pong handover	51	51	0
Learning-based method	*Count of service interruption*	3		
	Count of ping-pong handover	0	0	0

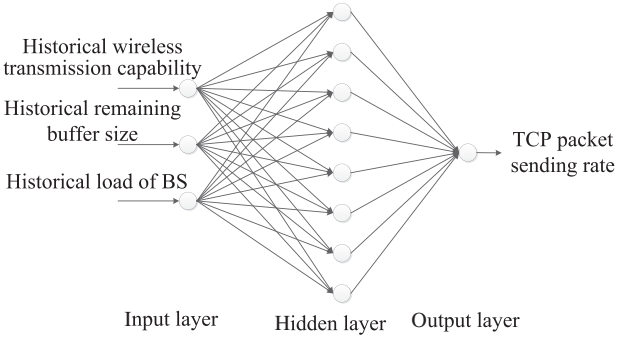

Figure 3.35 The structure of the BP neural network for TCP window optimization.

Big-Data-Assisted TCP Optimization

In current wireless networks, the TCP packet rate cannot be adjusted in a timely and accurate manner due to lack of knowledge on the real-time wireless transmission rate of the L2 layer. The traditional adjustment mechanism of the TCP window is based on trial and error [111, 112], which limits the overall system throughput and wastes wireless resources. With some useful RAN information, e.g., buffer size, load of the BS, the link throughput, service type, and PER, the TCP congestion window can be optimized and predicted to better match the radio channel variations. However, it is extremely hard to find a mathematical cross-layer model to determine the optimal TCP window with so many affecting factors. In this case, BD-assisted ML-based optimization offers an effective solution. With well constructed training data, a supervised learning-based model can be trained for the TCP window prediction.

A back propagation (BP) neural network can be utilized to predict the appropriate TCP window based on the context information of the wireless transmission rate. One example of the neural network is shown in Fig. 3.35. The inputs of the training model include the historic wireless transmission capability, historic remaining space in the buffer, and historic load of BS. The output is the rate at which the TCP should send out data packets. The prediction model was trained with 300,000 training data samples and 20,000 iterations.

The performance comparison of the traditional method and our proposed learning-based method is shown in Fig. 3.36. The x-axis shows the total simulation time of 12s,

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

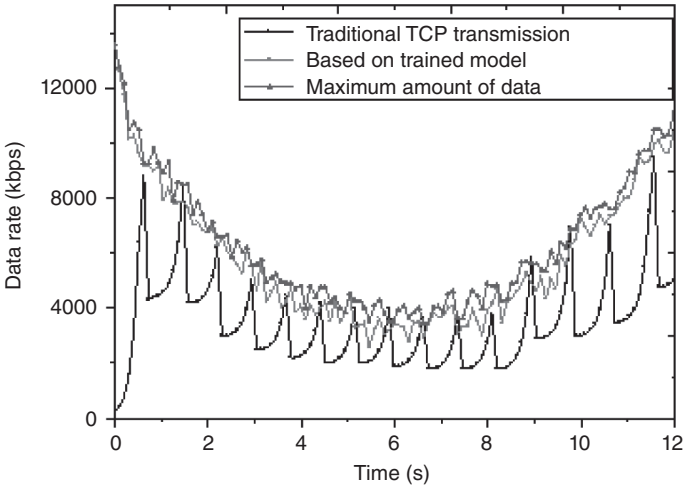

Figure 3.36 Comparison of packet sending rates.

and the y-axis is the data rate in the unit of kbps. The blue curve indicates the maximum amount of data that could be transmitted at each time instance. The black curve is the performance of the traditional TCP transmission scheme. The TCP packet sending rate initially increases to 2^{n+n_0} kbps, where n is the time index. It will drop to half when congestion is detected. Afterwards, it will rise with a rate of 2^{m+m_0}, where m is the time index starting from the last congestion point to the next congestion point. In the simulation, the update interval is 105ms; $n_0 = 8$; $m_0 = 3$. The red curve shows the achievable data rate of the BD-aided learning-based scheme. It can be seen that the BD-aided method is well-matched with the maximum transmission capability, while the traditional method shows great performance loss.

To measure the effectiveness of the BD-aided TCP window optimization, we calculate the similarity between the TCP packet sending rate vector and the maximum wireless transmission rate vector. The similarities for the traditional method and the proposed method are 0.867 and 0.953 respectively. A higher similarity value indicates better wireless resource utilization efficiency; thus the proposed method demonstrates a superior performance.

In summary, the BD-enabled learning-based method facilitates the dynamic optimization of TCP packet sending rate. It allows for a good match between the TCP window and the wireless channel condition. This significantly improves system throughput, buffer utilization, and the potential to reduce the retransmission.

Big-Data-Aided Beam Sweeping Optimization in Initial Access

In 5G NR, hybrid analog and digital beamforming is proposed in MIMO systems, which helps to balance performance and complexity, especially at high frequencies such as mmWave. In the initial access stage under the hybrid framework, beam sweeping

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

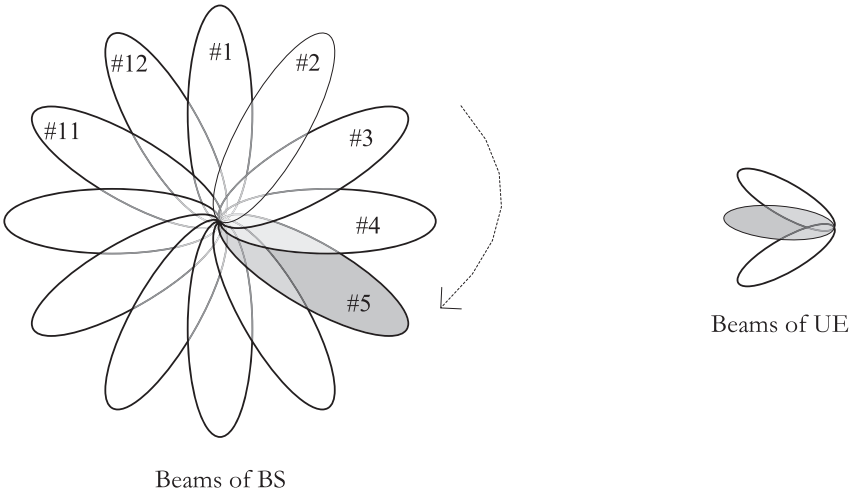

Figure 3.37 An example of beam sweeping.

is performed by the BS periodically to allow users to access the network from any direction in the cell. Specifically, the BS periodically transmits synchronization signals in different directions, covering the whole angular space. We assume that if the signal power received by the UE is greater than a threshold, the UE can successfully decode the signal and camp onto the network. As for high frequency communications, narrow beams are usually employed by the BS to compensate the overwhelmingly high path loss, and thus a large number of beam patterns may be needed to cover the whole angular space. If the UE also uses directional analog beams to receive synchronization signals, the exhaustive search for the desired beam pair may be prohibitively time-consuming.

At the conventional initial access stage, since there is no a prior shared information between the BS and the UE, they have to search for the desired beam pair sequentially. An example is shown in Fig. 3.37, wherein the BS transmits the synchronization signals in a beam-sweeping manner from beam #1 to beam #12. Imagine, when the UE roams to the cell, that the BS is transmitting the synchronization signals from beam #5 onwards, and then it has to wait for a long time until the synchronization signals are transmitted from the beam #4 direction again. As a result, the delay for initial access is large. Intuitively, the beam-sweeping process can be optimized if the users' spatial/angular distribution can be known at the BS.

Based on BD analytics, the BS is able to predict the user distribution by exploring the information feedback from UEs, and adjust the beam pattern and/or the beam-sweeping order according to the predicted user distribution. The information for predicting the user distribution includes the initial access delay of old UEs, the beam pair direction- and the received power of the synchronization signals. BD-aided initial access is suitable for the scenario where group characteristics exist among the users. For example, if the BS predicts that there is a high probability that a large number of users are distributed

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

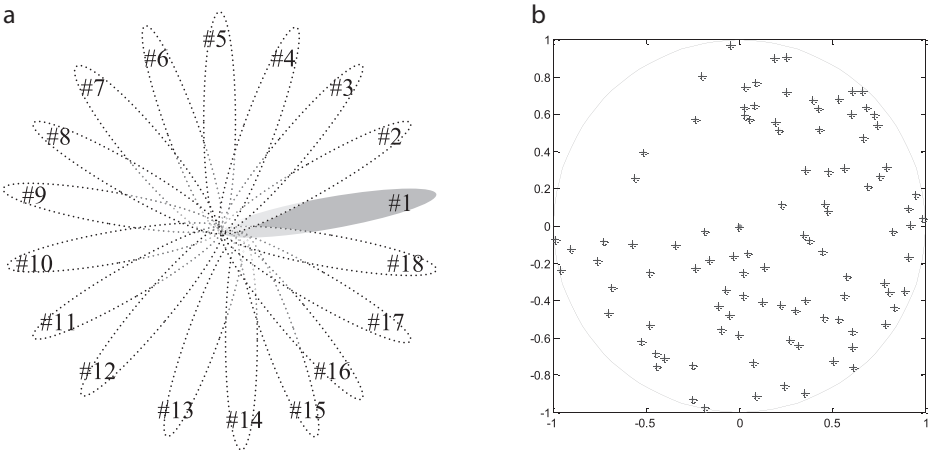

Figure 3.38 (a) Illustration of BS beams; (b) Illustration of user distribution.

in a certain direction, the beam-sweeping order should be adapted to guarantee that the users from these directions can discover the network in a timely manner. Therefore, the average initial access delay can be reduced.

In the following passage, some preliminary evaluation results are illustrated to show the effectiveness of the proposed BD-aided beam sweeping. In the simulation, we assume the BS has 16 beams to cover the whole cell shown in Fig. 3.38a, and the UE has one single beam direction. There are 100 users in the cell, and an example of the user distribution is shown in Fig. 3.38b. The BS has 64 antenna elements, and follows 8×8 uniform panel models with 20dB transmit power. The noise and cell radius is normalized to 1. The channel propagation is model as $h = d^{-3.7}$, where d is the distance between the UE and the BS. The access threshold is set as 0dB.

The following two schemes are evaluated for comparison:

- Conventional sequential beam sweeping: the BS sequentially scans the whole angular space from beam #1 to beam #16.
- BD-aided beam sweeping: The beam sweeping is optimized based on the information of UE access delay and access beam index, where high priority is allocated to the beam direction covering more users and larger access delay in the beam sweeping procedure.

In Fig. 3.39, the comparison of the average beam-sweeping times for the above two schemes are illustrated. It can be seen that the BD-aided scheme has much lower average beam-sweeping times. It indicates that the average access delay is significantly reduced. Besides, the proposed BD-aided scheme is quite robust in regards to the UE distribution. As we can see, with different degrees of user nonuniform distribution, the average beam-sweeping times at the BS are almost the same. The degree of the user nonuniform distribution is defined as the ratio of probability of the user located within the upper semicircle and the lower semicircle.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

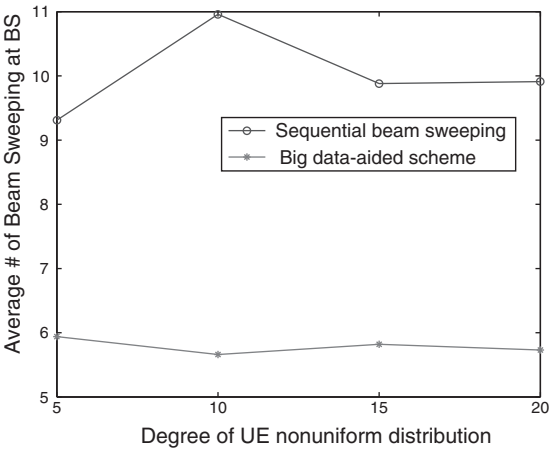

Figure 3.39 Comparison of initial access performance of the conventional scheme and the BD-aided scheme.

Summary

In this section, the benefits of applying WBD analytics to the wireless communication network are investigated. AI algorithms are first surveyed, with the basics briefly illustrated as examples. With BD analytics in the BD-enabled network architecture, e.g., in the CN or RAN or both, some potential performance improvements are expected in mobility management, cross layer TCP optimization, and beam-sweeping enhancement for initial network access. The user behavior, e.g., location and trajectory pattern, helps to improve the efficiency of the paging process in the idle state and the handover process in the connected state. Also, the BD-enabled learning-based method facilitates the dynamic optimization of TCP packet sending rate, achieving a good match between the TCP window and the wireless channel condition, significantly improving system throughput, buffer utilization and potential reduction of the retransmission numbers. Finally, the BD-aided beam-sweeping scheme, which is very important for MIMO operations at high frequency bands, shows much lower average beam-sweeping time. Accordingly, the average access delay is significantly reduced. Meanwhile, the proposed BD-aided scheme is robust against different UE distributions.

References

[1] ITU-R M.2083, "IMT vision-framework and overall objectives of the future development of IMT for 2020 and beyond," Sept. 2015.

[2] NGMN Alliance, "5G white paper," Feb. 2015.

[3] METIS 2020 Project, "The 5G future scenarios identified by METIS: The first step toward a 5G mobile and wireless communications system," Sept. 2013. www.metis2020.com/press-events/press/the-5g-future-scenarios-identified-by-metis/.

[4] M. Chiosi et al., "Network functions virtualisation: Introductory white paper," ETSI, Oct. 2012. www.etsi.org/technologiesclusters/technologies/nfv

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

[5] N. M. K. Chowdhury and R. Boutaba, "Network virtualization: State of the art and research challenges," *IEEE Comm. Mag.*, vol. 47, no. 7, pp. 20–26, 2009.

[6] B. Han, V. Gopalakrishnan, L. Ji, and S. Lee, "Network function virtualization: Challenges and opportunities for innovations," *IEEE Commun. Mag.*, vol. 53, no. 2, pp. 90–97, Feb. 2015.

[7] P. Veitch, M. J. McGrath, and V. Bayon, "An instrumentation and analytics framework for optimal and robust NFV deployment," *IEEE Commun. Mag.*, vol. 53, no. 2, pp. 126–133, Feb. 2015.

[8] H. Hawilo, A. Shami, M. Mirahmadi, and R. Asal, "NFV: State of the art challenges and implementation in next generation mobile networks (vEPC)," *IEEE Netw.*, vol. 28, no. 6, pp. 18–26, Nov. 2014.

[9] I. Silva, G. Mildh, and A. Kaloxylos, "Impact of network slicing on 5G radio access networks," *European Conference on Networks & Communications*, 2016.

[10] I. Silva, G. Mildh et al., "Tight integration of new 5G air interface and LTE to fulfill 5G requirements", *IEEE 81st Veh. Technol. Conf. (VTC Spring)*, Glasgow, May 2015.

[11] NGMN Alliance, "Description of network slicing concept," Version 1.0, Jan. 2016. www.ngmn.org/fileadmin/user_upload/160113_Network_Slicing_v1_0.pdf.

[12] 3GPP TR 38.801, "Study on new radio access technology: Radio access architecture and interfaces (Release 14)."

[13] 3GPP TS 23.501, "System architecture for the 5G system, (Release 15)."

[14] 3GPP TS 23.401, "General packet radio service (GPRS) enhancements for evolved universal terrestrial radio access network (E-UTRAN) access (Release 8)."

[15] 3GPP TS 38.401, "NG-RAN; Architecture description (Release 15)."

[16] 3GPP TS 38.300, "NR; Overall description; Stage-2 (Release 15)."

[17] 3GPP TS 37.340, "NR; Multi-connectivity; Overall description; Stage 2 (Release 15)."

[18] 3GPP TS 37.324, "Evolved universal terrestrial radio access (E-UTRA) and NR; Service data adaptation protocol (SDAP) specification (Release 15)."

[19] 3GPP TS 38.470, "NG-RAN; F1 general aspects and principles (Release 15)."

[20] 3GPP TR 36.933, "Study on context aware service delivery in RAN for LTE (Release 14)."

[21] 3GPP TR 36.300, "Evolved universal terrestrial radio access (E-UTRA) and evolved universal terrestrial radio access network (E-UTRAN); Overall description; Stage 2 (Release 8)."

[22] 3GPP TR 22.891, "Study on new services and markets technology enablers (Release 14)."

[23] 3GPP TR 38.913, "Study on scenarios and requirements for next generation access technologies (Release 14)."

[24] 3GPP TS 38.410, "NG-RAN; NG general aspects and principles (Release 15)."

[25] 3GPP TS 38.420, "NG-RAN; Xn general aspects and principles (Release 15)."

[26] ETSI, "Multi-access edge computing." www.etsi.org/technologies-clusters/technologies/multi-access-edge-computing.

[27] Open Networking Foundation, "Software-defined networking: The new norm for networks," Apr. 2012. http://www.opennetworking.org/images/stories/downloads/sdn-resources/white-papers/wp-sdn-newnorm.pdf

[28] J. Wan et al., "Software-defined industrial internet of things in the context of industry 4.0," *IEEE Sensors J.*, vol. 16, no. 20, pp. 7373–7380, 2016.

[29] S. Sezer et al., "Are we ready for SDN? Implementation challenges for software-defined networks," *IEEE Commun. Mag.*, vol. 51, no. 7, pp. 36–43, July 2013.

[30] B. Nunes et al., "A survey of software-defined networking: Past present and future of programmable networks," *IEEE Commun. Surveys & Tutorials*, vol. 16, pp. 1617–1634, 2014.

[31] The Open Group, "Service-oriented architecture standards." www.opengroup.org/standards/soa.

[32] C. L. I, C. Rowell, S. Han et al., "Toward green and soft: A 5G perspective," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 66–73, Feb. 2014.

[33] 3GPP TR 36.819, "Coordinated multi-point operation for LTE physical layer aspects (Release 11)."

[34] China Mobile Communications Corporation, "Simulation results for CoMP phase I evaluation in homogeneous network," 3GPP R1-111301, Meeting R1-65 contribution, Barcelona, May 2011

[35] Q. Wang, D. Jiang, G. Liu, and Z. Yan. "Coordinated multiple points transmission for LTE-advanced systems," in *Proc. of 5th Int. Conf. on Wireless Commun., Netw. and Mobile Comput., 2009 (WiCom '09)*, Sept. 2009, pp. 1–5.

[36] China Mobile Research Institute, "C-RAN white paper: The road towards green RAN," Jun. 2014.

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

[38] J. Wu, S. Rangan, and H. Zhang, *Green Communication*, CRC Press, 2013.

[39] NGMN, "Further study on critical C-RAN technologies," Apr., 2015.

[40] Next Generation Fronthaul Interface (1914) Working Group, 2015, http://sites.ieee.org/sagroups-1914.

[41] ETSI, "Network functions virtualisation," 2012. http://portal.etsi.org/portal/server.pt/community/NFV/367.

[42] B. A. A. Nunes, M. Mendonca, X.-N. Nguyen, K. Obraczka, and T. Turletti, "A survey of software-defined networking: Past, present, and future of programmable networks," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 3, pp. 1617–1634, Sept. 2014.

[43] C. L. I, Y. Yuan, J. Huang et al., "Rethink fronthaul for soft RAN", *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 82–88, 2015

[44] Y. Liu, and G. Liu, "User-centric wireless network for 5G," in *5G Mobile Communications*, W. Xiang, K. Zheng and X. Shen (eds.), Springer, 2016, pp. 457–474.

[45] 3GPP TS 36.300, "Evolved universal terrestrial radio access (E-UTRA) and evolved universal terrestrial radio access network (E-UTRAN); Overall description; Stage 2 (Release 8)."

[46] 3GPP TS 37.340, "NR; Multi-connectivity; Overall description; Stage-2 (Release 15)."

[47] 3GPP TS 36.410, "Evolved universal terrestrial radio access network (E-UTRAN) S1 general aspects and principles (Release 8)."

[48] 3GPP TS 36.413, "Evolved universal terrestrial radio access network (E-UTRAN) S1 application protocol (Release 11)."

[49] 3GPP TS 36.420, "Evolved universal terrestrial radio access network (E-UTRAN X2) general aspects and principles (Release 8)."

[50] 3GPP TS 36.423, "Evolved universal terrestrial radio access network (E-UTRAN) X2 application protocol (Release 8)."

[51] 3GPP TS 36.321, "Evolved universal terrestrial radio access (E-UTRA); medium access control (MAC) protocol specification (Release 8)."

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press

[52] 3GPP TS 36.211, "Evolved universal terrestrial radio access (E-UTRA); Physical channels and modulation (Release 8)."

[53] 3GPP TS 36.331, "Evolved universal terrestrial radio access (E-UTRA); Radio resource control (RRC); Protocol specification (Release 8)."

[54] 3GPP TS 36.212, "Evolved universal terrestrial radio access (E-UTRA); Multiplexing and channel coding (Release 8)."

[55] 3GPP TS 36.322, "Evolved universal terrestrial radio access (E-UTRA); Radio link control (RLC) protocol specification (Release 8)."

[56] 3GPP TS 36.323, "Evolved universal terrestrial radio access (E-UTRA); Packet Data convergence protocol (PDCP) specification (Release 8)."

[57] 3GPP TR 38.804, "Study on new radio access technology radio interface protocol aspects (Release 14)."

[58] X. Cheng et al. "Mobile big data: The fuel for data-driven wireless," *IEEE Internet of Things J.*, vol. 4, no. 5, pp. 1489–1516, Oct. 2017. DOI 10.1109/JIOT.2017.2714189.

[59] S. Han, C.-L. I, S. Wang et al., "Big data enabled mobile network design for 5G and beyond," accepted by *IEEE Commun. Mag.*, 2017.

[60] J. Qadir et al., "Artificial intelligence enabled networking," *IEEE Access*, vol. 3, pp. 3079–3082, 2015.

[61] X. Wang, X. Li, and V. C. M. Leung, "Artificial intelligence-based techniques for emerging heterogeneous network: State of the arts, opportunities, and challenges,", *IEEE Access*, vol. 3, pp. 1379–1391, 2015.

[62] M. Bkassiny, Y. Li, S. K. Jayaweera et al., "A survey on machine-learning techniques in cognitive radios," *IEEE Commun. Surveys and Tutorials*, vol. 15, no. 3, pp. 1136–1159, 2013.

[63] M. A. Alsheikh, S. Lin , D. Niyato et al., "Machine learning in wireless sensor networks: Algorithms, strategies, and applications," *IEEE Commun. Surveys and Tutorials*, vol. 16, no. 4, pp. 1996–2018, 2014.

[64] C. Jiang, H. Zhang, Y. Ren et al., "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 98–105, Apr. 2017.

[65] R. Li, L. Zhao X. Zhou, et al., "Intelligent 5G: When cellular networks meet artificial intelligence," *IEEE Wireless Commun.*, vol. 24, no. 5, pp. 175–183, Oct. 2017.

[66] P. V. Klaine, A. I. Muhammad, O. Oluwakayode et al., "A survey of machine learning techniques applied to self organizing cellular networks," *IEEE Commun. Surveys & Tutorials*, vol.19, no.7, pp. 2392–2431, fourth quarter, 2017.

[67] S. Hu, Y.-d. Yao, and Z. Yang, "MAC protocol identification using support vector machines for cognitive radio networks," *IEEE Wireless Commun.*, vol. 21, no. 1, pp. 52–60, Feb. 2014.

[68] R. Li, L. Zhao, X. Chen et al., "TACT: A transfer actor-critic learning framework for energy saving in cellular radio access networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, pp. 2000–2011, Apr. 2014.

[69] N. Sinclair, D. Harle, I. A. Glover, J. Irvine, and R. C. Atkinson, "An advanced SOM algorithm applied to handover management within LTE," *IEEE Trans. on Veh. Technol.*, vol. 62, no. 5, pp. 1883–1894, Jun. 2013.

[70] P. Wang, S. C. Lin, and M. Luo, "A framework for QoS-aware traffic classification using semi-supervised machine learning in SDNs," *IEEE Int. Conf. on Services Comput. (SCC)*, San Francisco, pp. 760-765, 2016.

[71] P. Amaral, J. Dinis, P. Pinto et al., "Machine learning in software defined networks: Data collection and traffic classification," *IEEE 24th Int. Conf. on Netw. Protocols (ICNP)*, Singapore, pp. 1–5, 2016.

[72] I. Yahia, J. Bendriss, A. Samba et al., "CogNitive 5G networks: Comprehensive operator use cases with machine learning for management operations," *20th Conf. on Innovations in Clouds, Internet and Netw. (ICIN)*, Paris, pp. 252–259, 2017.

[73] X. Gao, L. Dai, Y. Sun et al., "Machine learning inspired energy-efficient hybrid precoding for mmWave massive MIMO systems," *IEEE Int. Conf. on Commun. (ICC)*, 2017.

[74] J. Joung, "Machine learning-based antenna selection in wireless communications," *IEEE Commun. Lett.*, vol. 20, no. 11, pp. 2241–2244, Nov. 2016.

[75] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2007.

[76] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, The MIT Press, 2012.

[77] E. Alpaydin, *Introduction to Machine Learning*, Second Edition, The MIT Press, 2010.

[78] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition, Springer, 2016.

[79] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, Second Edition, The MIT Press, 2018.

[80] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, The MIT Press, 2016.

[81] N. Kato et al., "The deep learning vision for heterogeneous network traffic control: Proposal, challenges, and future perspective," *IEEE Wireless Commun.*, vol. 24, no. 3, pp. 146–153, 2017.

[82] D. Liu, B. Chen, C. Yang, and A. Molisch, "Caching at the wireless edge: Design aspects, challenges and future direction," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 22–28, Sept. 2016.

[83] Wikipedia, "Deep learning." https://en.wikipedia.org/wiki/Deep_learning.

[84] C. Yao, C. Yang, Z. Xiong, "Energy-saving predictive resource planning and allocation," *IEEE Trans. on Commun.*, vol. 64, no. 12, pp. 5078–5095, Dec. 2016.

[85] S. Zhou, D Lee, B Leng, et al., "On the spatial distribution of base stations and its relation to the traffic density in cellular networks," *IEEE Access*, vol. 3, pp. 998–1010, 2015.

[86] C.-L. I, Y. Liu, S. Han, S. Wang, and G. Liu, "On big data analytics for greener and softer RAN," *IEEE Access*, vol. 3, pp. 3068–3075, Mar. 2015.

[87] 3GPP SA2, "Study of enablers for network automation for 5G," Doc number: S2-173827, May. 2017, Hangzhou, China.

[88] X. Zhang et al., "Social computing for mobile big data," *Computer*, vol. 49, no. 9, pp. 86–90, Sept. 2016.

[89] X. Cheng, L. Fang, X. Hong, and L. Yang, "Exploiting mobile big data: Sources, features, and applications," *IEEE Netw.*, vol. 31, no. 1, pp. 72–79, Jan./Feb. 2017.

[90] S. Bi, R. Zhang, Z. Ding, and S. Cui, "Wireless communications in the era of big data," *IEEE Commun. Mag.*, vol. 53, no. 10, pp. 190–199, Oct. 2015.

[91] K. Zheng, Z. Yang, K. Zhang et al., "Big data-driven optimization for mobile networks toward 5G," *IEEE Netw.*, vol. 30, no. 1, pp. 44–51, Jan. 2016.

[92] C.-L. I, Q. Sun, Z. Liu, S. Zhang, and S. Han, "The big-dat-driven intelligent wireless network: Architecture, use cases, solutions, and future trends," *IEEE Veh. Technol. Mag.*, vol. 12, no. 4, pp. 20–29, Dec., 2017.

[93] FuTURE Forum 5G SIG, Whitepaper, "Wireless big data for smart 5G," Nov. 2017, available: http://www.future-forum.org/dl/171114/whitepaper2017.rar.

[94] Y. Li, B. Cao, and C. Wang, "Handover schemes in heterogeneous LTE networks: Challenges and opportunities," *IEEE Wireless Commun.*, vol. 23, no. 2, pp. 112–117, Feb. 2016.
[95] 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. 474–482, Mar. 2016.
[96] A. Ray, S. Deb, and P. Monogioudis, "Localization of LTE measurement records with missing information," *IEEE INFOCOM 2016*.
[97] J. Tadrous, A. Eryilmaz, and H. E. Gamal, "Proactive content download and user demand shaping for data networks," *IEEE/ACM Transactions on Networking*, vol. 23, no. 6, pp. 1917–1930, 2015.
[98] E. Baştuğ, M. Bennis, M. Kountouris, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," *EURASIP Journal on Wireless Communications and Networking*, vol. 2015, no. 1, pp. 1–11, 2015.
[99] D. Liu and C. Yang, "Cache-enabled heterogeneous cellular networks: Comparison and tradeoffs," IEEE ICC, 2016, accepted.
[100] D. Liu and C. Yang, "Energy efficiency of downlink networks with caching at base stations," Accepted, IEEE Journal on Selected Areas in Communications, 2015. http://arxiv.org/abs/1505.06615.
[101] H. Li and G. Ascheid, "Long-term window scheduling in multiuser OFDM systems based on large scale fading maps," IEEE SPAWC 2012 *IEEE SPAWC 2012*.
[102] Z. Lu and G. De Veciana, "Optimizing stored video delivery for mobile networks: The value of knowing the future," IEEE INFOCOM 2013 *IEEE INFOCOM 2013*.
[103] J. Tadrous, A. Eryilmaz, and H. El Gamal, "Proactive resource allocation: Harnessing the diversity and multicast gains," *IEEE Trans. Information Theory*, vol. 59, no. 8, pp. 4833–4854, 2013.
[104] C. Yao, B. Chen, and C. Yang, "Energy Saving pushing based on personal interest and context information," Accepted, *Proc. IEEE Vehicular Technology Conference 2016*.
[105] V. A. Siris and D. Kalyvas, "Enhancing mobile data offloading with mobility prediction and prefetching," ACM MobiArch, 2012.
[106] S.-Y. Lien, K.-C. Chen, and Y.-C. Liang, "Ultra-low latency ubiquitous connections in heterogeneous cloud radio access networks," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 22–31, Jun. 2015.
[107] Y. Huang, J. Tan, and Y.-C. Liang, "Wireless big data: Transforming heterogeneous networks to smart networks," *Journal of Communications and Information Networks*, vol. 2, no. 1, pp. 19–32, Mar. 2017.
[108] Y. Sun, G. Feng, S. Qin, S. Sun, and L. Zhang, "User behavior aware cell association in heterogeneous cellular networks," *Proc. of IEEE Wireless Communications and Networking Conference (WCNC)*, San Francisco, Mar. 2017.
[109] 3GPP TS 36.331, "Radio resource control (RRC); Protocol specification (Release 14)."
[110] N. Sinclair, D. Harle, I. A. Glover, J. Irvine, and R. C. Atkinson, "Parameter optimization for LTE handover using an advanced SOM algorithm," *Proc. IEEE Vehicular Technology Conference 2013*.
[111] W. Stevens, "TCP slow start, congestion avoidance, fast retransmit, and fast recovery algorithms." Request for comment: 2001, 1997
[112] S. Bajeja and A. Gosai, "Performance evaluation of traditional TCP variants in wireless multihop networks," *3rd Int. Conf. on Comput. for Sustainable Global Development (INDIACom)*, New Delhi, India, pp. 3517–3522, 2016.

https://doi.org/10.1017/9781108277389.004 Published online by Cambridge University Press