

Chapter 1

Future Cellular Network Architecture

Ying Li

Samsung Research America

CONTENTS

1.1	Introduction	4
1.2	Today's Cellular Network Architecture	6
1.2.1	Radio Access Network (RAN)	6
1.2.2	Evolved Packet Core (EPC)	8
1.3	Future Radio Access Networks	9
1.3.1	UE's Heterogeneous Traffic to Heterogeneous eNBs	10
1.3.2	Cloud-RAN	12
1.3.3	Adaptive and Self-Organized RAN with Drop-and-Play Small Cells	14
1.4	Future Evolved Core Network	17
1.4.1	Mobile SDN	17
1.4.2	Network Virtualization in EPC	18
1.5	Conclusion	20
	References	20

1.1 Introduction

The demand of wireless data traffic is explosively increasing [26] due to the increasing popularity of smart phones and other mobile data devices such as tablets, netbooks, and eBook readers among consumers and businesses. The fourth generation (4G) cellular technologies [15] including Long Term Evolution (LTE)-Advanced and Advanced Mobile WiMAX (IEEE 802.16m) use traditional network architecture, which has limitations. In order to meet this spectacular growth in mobile data traffic, improvements in future cellular network architecture has been of paramount importance.

Today's cellular networks consist of Radio Access Network (RAN), which mainly deals with the air interface of the base stations (referred to as evolved Node Bs (eNBs)) and mobile stations (referred to as user equipment (UE)), where an eNB can consist of one or multiple cells, and Evolved Packet Core (EPC) network, which mainly deals with the packet processing after the eNB before it goes to the Internet [24]. The explosive increase in demand for wireless data has placed increasing challenges on today's RAN and EPC, which both have limitations.

In RAN, a possibility to increase the overall system capacity is to deploy a large number of smaller cells. In today's RAN, compared to traditional homogeneous networks, there are new scenarios and considerations in heterogeneous networks with base stations of diverse sizes and types [18, 27]. One consideration is traffic offloading from large cells to small cells. For example, small cell's footprint can be enlarged to offload the traffic from a macro cell to a small cell [2]. Another consideration is on the resource management considering interferences. The large cells and small cells can be deployed on multiple carriers, and carrier aggregation (CA) [24] can be used to achieve high system performance. The large cells and small cells can be deployed on a single carrier and co-channel intercell interference management can apply, such as time domain muting. Coordinated Multipoint (CoMP) transmissions [5] can also be used to coordinate the transmissions among multiple cells, such as a joint transmission from multiple points to achieve higher system performance.

However, even with the techniques above for heterogeneous networks, today's RAN is with limitations and is still not optimized in several aspects.

1. First, even though CA or CoMP can be used for multiple cells with different sizes, the current RAN can only support the case that the multiple cells involved are with a same eNB. An eNB can have multiple cells, such as a large cell and a few small cells by remote radio head (RRH), transmit point (TP), etc. A CA or CoMP involving more than one eNB currently is not supported. In addition, the current CA does not much support the deployment scenario where the cells are connected via nonideal backhaul (meaning the backhaul delay is nonnegligible). These limit the deployment scenarios as well as the system performance.
2. Second, the operational expenditure (OPEX) of today's RAN is high. For example, RAN nodes constitute most of the energy consumption of a cellular network [11]. The large number of RAN nodes are usually based on proprietary platforms. A RAN node utilization is usually lower than that capacity

because the system is designed to cover the peak load; however, the average load is far lower, but today's support for resource sharing among RAN nodes is low. The high OPEX makes it difficult for operators to increase the revenue while the mobile traffic is explosively increasing.

3. Third, the support for self-organized networking (SON) is limited. Some SON functions are supported [24], such as plug-and-play as a home eNB. With the advances in wireless backhaul and energy harvesting, the future small cells can be with no wire (no wired backhaul or powerline), hence, they can be drop-and-play. Today's SON functions are limited and not ready for possible future drop-and-play scenarios.

The future RAN should improve to mitigate the limitations. The following can be considered for the future RAN:

1. A better support for a UE to be connected to multiple eNBs concurrently is needed, with a consideration of diverse backhaul conditions. In addition, diverse applications and diverse traffic of a UE can be assigned to cells with diverse backhaul conditions concurrently, to further improve the system performance.
2. Technologies to reduce the OPEX are needed. One of the potential technologies is cloud RAN [22], where the baseband processing of different eNBs and cells can be centralized and resources pooled and shared. It can leverage more efficient resource utilization among different eNBs and cells.
3. Advanced SON to support drop-and-play small cells can be considered. The advanced SON should support load balancing, robust routes establishment and adaptive routing, adaptive ON/OFF of small cells considering energy harvesting and traffic dynamics, and so on.

For EPC, today's architectures have some major limitations. A centralized data plane in the cellular network forces all the traffic of the UEs (including traffic between users on the same cellular network) to go through the packet gateway (P-GW) at the cellular-Internet boundary, which faces scalability challenges and makes it difficult to host popular content inside the cellular network. In addition, the network equipment has vendor-specific configuration interfaces and communicates through complex control plane protocols, with a large and growing number of tunable parameters (e.g., several thousand parameters for eNBs). As such, network operators have (at best) indirect control over the operation of their networks, with little ability to create innovative services.

Network operators are finding it difficult to introduce new revenue generating services and optimize their expensive infrastructures. Networks continue to have serious known problems with security, robustness, manageability, and mobility. Network capital costs have not been reducing fast enough, and operational costs have been growing. Even vendors and third parties are not able to provide customized cost-effective solutions to address their customers' problems.

The limitations in EPC have created opportunities for the next generation wireless network architectures incorporating Software Defined Networking (SDN) and network virtualization. SDN and virtualization are gaining momentum in wired networks [19–21] because of their advantages, such as low-cost deployments, easy management, and so on. However, SDN and virtualization have not been studied much for wireless networks, yet they have great potential to meet the challenges that today's wireless networks are facing [16, 17, 23].

SDN is a type of networking in which the control plane is physically separate from the forwarding plane. Network intelligence is (logically) centralized in a software-based SDN controller, which maintains a global view of the network. The network appears to the application and policy engines as a single, logical switch. As for network virtualization, the resources and functions in the cellular network can be virtualized. As such, SDN and network virtualization can adaptively and flexibly provide traffic offloading, revenue-adding services, and the capability to deliver services to mobile stations that require a large amount of data.

The concept and technologies of SDN and virtualization for cellular networks are still in a very early stage. They have to address cellular networks' own unique characteristics and requirements. For example, special considerations are needed for aspects such as supporting many subscribers, frequent mobility, fine-grained measurement and control, and real-time adaptation, which introduces scalability challenges.

Accordingly, the chapter is organized as follows. In the next section we describe today's cellular network architecture, including RAN and EPC. The third section discusses the future architecture for RAN, including the support for a UE to connect concurrently to multiple eNBs and associate diverse traffic to the eNBs, Cloud-RAN, and advanced SON to support drop-and-play small cells. In the fourth section, we provide the future architecture for EPC, including SDN and network virtualization. The chapter is concluded with some summarizing visions.

1.2 Today's Cellular Network Architecture

Today's cellular network architecture is discussed in this section. Figure 1.1 illustrates an overview of the architecture of RAN and EPC. RAN and EPC are interfaced via S-interface, mostly S1-interface [24]. The details of the figure are presented in the following subsections.

1.2.1 Radio Access Network (RAN)

Today's RAN can deploy heterogeneous networks. Cells with different sizes can be used in a hierarchical network deployment, referred to as multitier deployment or multitier networks, where each tier can be for one type of cell of certain size. The type and location of the eNB controlling these cells will play a significant role in determining the cost and performance of the multitier deployments. For example, indoor femtocell deployments using home eNBs (HeNBs) can utilize the existing

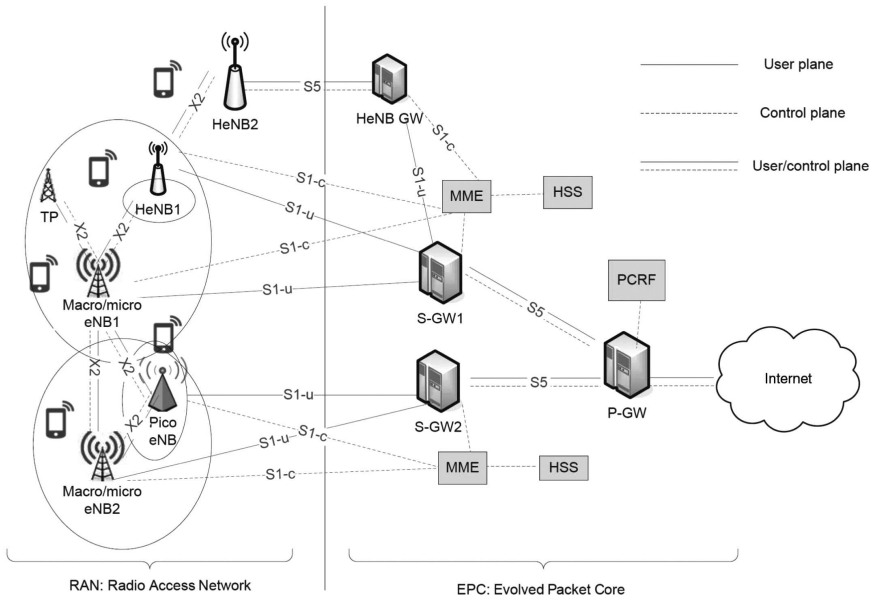


Figure 1.1: An overview of today's cellular network architecture.

backhaul, thereby significantly lowering the cost of such deployments. With outdoor picocell deployments through pico eNB, the operator will need to provide backhaul capability and manage more critical spectrum reuse challenges. Other deployment models cover indoor enterprise or outdoor campus deployments that may impose different manageability and reliability requirements.

The RAN part of Figure 1.1 illustrates an exemplary heterogeneous network with macro/micro eNB, pico eNB, and a femtocell/HeNB. For the heterogeneous network shown in Figure 1.1 the pico eNB has smaller transmission power than macro eNB—hence with smaller coverage than the macro eNB. On the other hand, the HeNB can have smaller transmission power than a pico eNB. A cell formed by a TP can belong to an eNB. Picocells typically are managed together with macro/microcells by operators. An interface of X2 can be used for the communications among the eNBs, HeNBs, and TPs. All kinds of eNBs can be connected to the servicing-gateway (S-GW) in EPC for the user plane (or the data plane), and connected to Mobility Management Entity (MME) in EPC for the control plane. For HeNB, it can also be connected to the S-GW and MME via a HeNB gateway. It is noted that Figure 1.1 does not include relay eNB for simplicity. Relay eNB can be included where an eNB that does not have wired backhaul can connect to the EPC via a relay eNB.

The coverage area of the picocell is limited not only by its transmit power, but also, to a large extent, by the intercell interference from other cells. Therefore, if the cell selection criteria are only based on downlink UE measurements such as the reference signal received power (RSRP), only UEs in close vicinity will end up being

served by the pico eNB. Due to the higher deployment density of the small cells, it is beneficial to expand the footprint of the picocells, i.e., offloading UEs from macrocells to picocells, to enable more UEs to connect to the small cells to take advantage of the higher deployment density. This can be achieved through cell range expansion (RE) [2]. One of the approaches for cell range expansion is that a cell-specific bias to the UE measurement of X dB is applied for pico eNB to favor connecting to it. In this way, more UEs will be inclined to connect to pico eNBs instead of macro eNBs. Furthermore, time domain intercell interference coordination techniques can also be utilized for pico users that are served at the edge of the serving pico cell, for example, for traffic offloading from a macrocell to a picocell.

Spectrum allocation across multiple tiers is an important aspect of deployment and use of hierarchical architectures. According to the spectrum used, multitier cell deployments are possible for the following cases:

1. Multiple-carrier case: The multitier cells are deployed on multiple carriers. When multiple carriers are available, choices can be made to enable flexible cell deployment. For example, the macrocell and small cells can be deployed on distinct carriers, or on the same set of carriers while having joint carrier and power assignment/selection to better manage intercell interference.
2. Single-carrier case: The multitier cells are deployed on a single carrier. This can also be called co-channel deployment.

Techniques such as CA and CoMP can apply for the resource management. Muting in the time domain can also apply [3].

A UE can be connected to multiple cells, such as in the CA case or CoMP case. However, today's RAN does not support a UE concurrently connecting to more than one eNB. It does not support much for CA in the scenarios where cells can be connected via nonideal backhaul links.

1.2.2 Evolved Packet Core (EPC)

Today's cellular networks connect eNBs to the Internet using IP networking equipment. An illustration of the entities in EPC and how they connect to each other is shown in the EPC part of Figure 1.1.

For the data plane or the user plane, the traffic from an eNB goes through a serving gateWay (S-GW) over a tunnel. The S-GW serves as a local mobility anchor that enables seamless communication when the user moves from one base station to another. The S-GW must handle frequent changes in a user's location, and store a large amount of states since users retain their IP addresses when they move. The S-GW tunnels traffic to the P-GW. The P-GW enforces quality of service (QoS) policies and monitors traffic to perform billing. The P-GW also connects to the Internet and other cellular data networks, and acts as a firewall that blocks unwanted traffic. The policies at the P-GW can be very fine grained, based on whether the user is roaming, properties of the user equipment, usage caps in the service contract, parental controls, and so on.

Besides data plane functionalities, the eNB, S-GW, and P-GW also participate in several control plane protocols. In coordination with the MME, they perform hop-by-hop signaling to handle session setup, teardown, band reconfiguration, as well as mobility, e.g., location update, paging, and handoff. For example, in response to a UE's request for dedicated session setup (e.g., for VoIP call), the P-GW sends QoS and other session information (e.g., the TCP/IP 5-tuple) to the S-GW. The S-GW in turn forwards the information to the MME. The MME then asks the eNB to allocate radio resources and establish the connection to the UE. During handoff of a UE, the source eNB sends the handoff request to the target eNB. After receiving an acknowledgment, the source eNB transfers the UE state (e.g., buffered packets) to the target eNB. The target eNB also informs the MME that the UE has changed cells, and the previous eNB to release resources.

The S-GW and P-GW are also involved in routing protocols. The Policy Control and Charging Function (PCRF) manages flow-based charging in the P-GW. The PCRF is connected to the P-GW via the control interface. The PCRF also provides the QoS authorization (QoS class identifier and bit rates) that decides how to treat each traffic flow, based on the user's subscription profile. QoS policies can be dynamic, e.g., based on time of day. This must be enforced at the P-GW. The Home Subscriber Server (HSS) contains subscription information for each user, such as the QoS profile, any access restrictions for roaming, and the associated MME. The HSS is connected to the MME via the control interface. In times of cell congestion, a base station reduces the max rate allowed for subscribers according to their profiles, in coordination with the P-GW.

Today's EPC has limitations. Centralizing data plane functions such as monitoring, access control, and quality of service functionality at the P-GW introduces scalability challenges. This makes the equipment very expensive (e.g., more than \$6 million for a Cisco P-GW). Centralizing data plane functions at the cellular-Internet boundary forces all traffic through the P-GW, including traffic between users on the same cellular network, making it difficult to host popular content inside the cellular network. In addition, the network equipment has vendor-specific configuration interfaces, and communicates through complex control plane protocols, with a large and growing number of tunable parameters (e.g., several thousand parameters for base stations). As such, carriers have (at best) indirect control over the operation of their networks, with little ability to create innovative services.

1.3 Future Radio Access Networks

Future architecture for RAN is discussed in this section, including the support for a UE to connect concurrently to multiple eNBs and associate diverse traffic to the eNBs, Cloud-RAN, and advanced SON to support drop-and-play small cells.

1.3.1 UE's Heterogeneous Traffic to Heterogeneous eNBs

As mentioned in Section 1.2.1, today's CA or CoMP technology is limited to the case that a UE is associated with multiple cells within an eNB. Such limitation has disadvantages, such as the resources among multiple eNBs cannot enjoy the advantage that CA or CoMP can bring, such as improved spectrum efficiency, enhanced radio resources sharing, etc. Hence, it is desirable to extend CA or CoMP technologies to the case that a UE can be associated to multiple eNBs. For the CA case, today's CA does not support much for the scenarios that the cells are connected via nonideal backhaul. The technology to allow in a future RAN that a UE can be connected to multiple eNBs where eNBs can be connected via ideal or nonideal backhaul is currently discussed in 3GPP [1, 7], and such technology is referred to as dual connectivity.

The dual-connectivity technology includes multistream aggregation operation. The multistream operation involves dual UE connectivity to two eNBs that are not necessarily collocated and may not have ideal backhaul connection. Enabling dual connectivity for FDD-only and TDD-only systems is discussed in [1]. All technical reasons for endorsing dual connectivity (e.g., mobility robustness and better throughput performance) for FDD-only and TDD-only systems are equally applicable for TDD-FDD multistream aggregation. Enabling dual connectivity for TDD-FDD systems discussed in [4, 10]. Examples of benefits would mainly be increased mobility robustness and increased throughput.

Figure 1.2 shows an exemplary dual-connectivity scenario. In the figure, a UE can be connected to a large eNB, and a small eNB, where the large eNB may not be connected to the small eNB via ideal backhaul; i.e., the backhaul can have nonnegligible delay.

As a further consideration, when a UE can be connected to multiple eNBs and cells where eNBs and cells can be connected via ideal or nonideal backhaul, diverse applications and diverse types of traffic of a UE can be associated to eNBs and cells with diverse backhaul conditions concurrently, to further improve the system performance. More details in this aspect are discussed below.

The future UEs may see many heterogeneous cells with different loads, wireless/wired backhaul conditions on rate, delays, etc., and the radio access links in-between a UE and cells can be diverse in rate and delay. Meanwhile, a UE can have various concurrent traffic flows with diverse QoS requirements, such as delay stringent interactive video and delay astringent best effort data.

One of the important problems in cellular networks is to properly associate UEs with serving cells. This problem usually is referred to as the user association problem. The user association would impact the radio resource allocation to UEs, and hence impact the QoS. For example, to guarantee QoS, delay stringent traffic of a UE can be associated to a first eNB via which the total delay of the UE to the eNB and the eNB to the EPC is relatively small, while delay astringent traffic of a UE can be associated to a second eNB via which the total delay of the UE to the eNB and the eNB to the EPC is relatively large. It is of great interest to study in the future cellular network how to associate UEs with eNBs and cells, given the emerging new

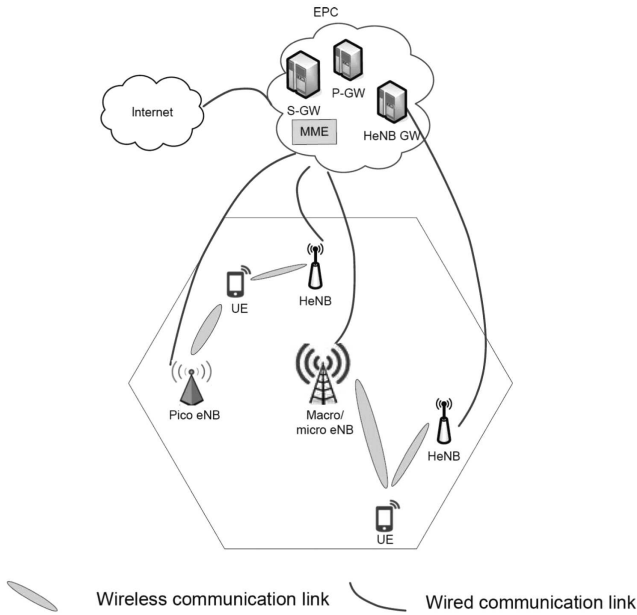


Figure 1.2: An exemplary dual-connectivity scenario.

scenarios that the UEs can have diverse application traffic flows and the cells can have diverse conditions and status.

Traditionally in cellular networks, the simplest rule is to choose the eNB that gives the strongest downlink reference signal. However, such a rule has limitations in the sense that it does not consider other factors such as cell load balancing. There have been efforts in the literature toward developing user association rules considering cell load balancing; for example, a cell can broadcast its load and a UE can be associated with a cell with lower load [13]. For a heterogeneous network, the cell association has further consideration. Small cell range expansion can expand the footprint of small cells that can coexist with large cells [2], as mentioned in Section 1.2.1. All of the above have been based on a UE being associated with only one eNB; however, concurrent connection where a UE can be connected to multiple eNBs may provide more freedom and benefits considering diverse concurrent traffic of a UE.

Moreover, diverse backhaul conditions of the eNBs and cells are playing another role for the consideration of the UE cell association rule. The impact on different back-haul conditions on CoMP performances has been studied [5], mainly in the aspect of how the backhaul delay could affect the coordinated joint scheduling for multiple transmission points. A new study item in 3GPP standardization for small cell enhancement also mentions the further study on backhaul impact on the performance [8]. However, it is not well studied how the backhaul conditions would impact the UE's diverse traffic associating with the diverse eNBs and cells. Not considering

the eNB backhaul may not be effective in resource allocation. For instance, if a UE has a strong wireless access link to an eNB but the eNB has a backhaul with large latency, if the cell association is decided by wireless access, the UE would be associated to this eNB; however, considering the large delay at the eNB's backhaul, the time stringent traffic of the UE may not be satisfied with the total delay.

In [29], the problem of assigning UE traffic flows concurrently to multiple BSs is studied, so that each flow's QoS requirement is satisfied, with the eNB backhaul condition taken into account. It shows that the dual connectivity can bring the benefits of UE-satisfied QoS for diverse traffic types. It proposes a framework to associate traffic flows with different QoS requirements of a UE to multiple eNBs, so that the corresponding QoS requirements can be satisfied, where the heterogeneous backhaul conditions of eNBs are taken into account.

For future RAN, further studies are needed to enable a UE with diverse traffic to be connected to multiple eNBs with diverse backhaul conditions.

1.3.2 *Cloud-RAN*

Mobile network traffic is significantly increasing by the demand generated by application of mobile devices, while the revenue is difficult to increase. To keep profit, mobile operators must reduce OPEX as well as continuously develop and provide better services to their customers.

One consideration is to reduce the operational cost by sharing resources among eNBs. A RAN node utilization is usually lower than that capacity because the system is designed to cover the peak load. The network should adequately support the peak load; however, it brings higher cost due to good QoS provisioning for the peak load. Possible pooling of eNBs may reduce the cost by sharing the resource among eNBs for more efficient resource utilization. It can also reduce power consumption to reduce the total cost of ownership (TCO), which is very important as the electricity bill takes a large portion of the TCO (about 20% according to [11]) and the cell sites are the major source of the energy consumption (about 70% according to [11]).

Cloud-RAN (C-RAN) technology [22] uses resource pooling and virtualization of eNBs. C-RAN stands for centralized processing, collaborative radio, real-time cloud computing, clean RAN system. It can leverage more efficient resource utilization among eNBs. It has great potential to help reduce OPEX. C-RAN is to realize RAN nodes onto standard IT servers, storages, and switches bringing advantages, such as lower footprint and energy consumption coming from dynamic resource allocation and traffic load balancing, easier management and operation, and faster time to market. In major mobile operators' networks, multiple RAN nodes from multiple vendors are usually operated with different mobile network systems, e.g., 3G and LTE, in the same area. These multiple platforms expect to be consolidated into a physical eNB based on IT virtualization technologies, referred to as eNB virtualization.

Figure 1.3 illustrates an exemplary C-RAN architecture. In the figure, the distributed radio unit (RU) of eNBs and TPs in a same area can be connected to a digital unit (DU) or DU cloud, which is a group of DUs, via a high-bandwidth and

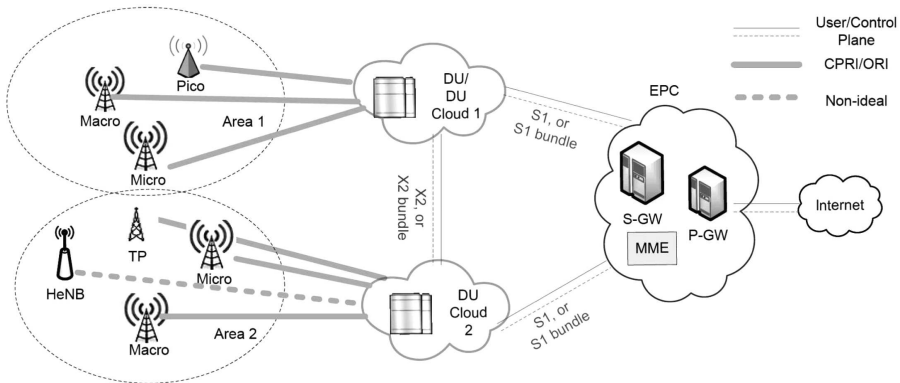


Figure 1.3: An exemplary C-RAN architecture.

low-latency transport network Common Public Radio Interface (CPRI) or Open Radio Interface (ORI). The DU can be also referred to as baseband unit (BBU). The DU and DU cloud of an area are centralized in one physical location for providing resource aggregation and pooling. The DU and DU cloud include the radio functions of the digital baseband domain. The DU is in charge of the channel coding, digital signal processing (DSP), modulation/demodulation process, and interface module. The RU is different from the traditional eNB or TP, as its baseband processing is moved to DU. The DU and DU cloud are connected to each other via an X2 or X2 bundle, and they are connected to the EPC via an S1 or S1 bundle.

ENB virtualization requires baseband radio processing using IT virtualization technologies, such as high-performance general purpose processors and real-time processing virtualization to provide required signal processing capacity. eNB virtualization for C-RAN moreover requires building the processing resource, i.e., DU, pool for aggregating the resources onto a centralized virtualized environment, such as cloud infrastructure.

To support CoMP transmission/reception, UE data and channel information need to be shared among eNBs/DUs, and a high-bandwidth and low-latency interconnection for real-time cooperation among these should be supported on the virtualized environment.

To further reduce the TCO for mobile operators, SON can be used to support C-RAN. SON is especially useful as the number and structure of network parameters have become large and complex, quick evolution of wireless networks has led to parallel operation of 2G, 3G, and EPC infrastructures, and the rapidly expanding number of eNBs needs to be configured and managed with the least possible human interaction. An effective SON solution for C-RAN must be multivendor by nature and leverage the timely information from protocols such as X2 for better handling of CoMP and for such routine mobile network functions as handover and mobility optimizations.

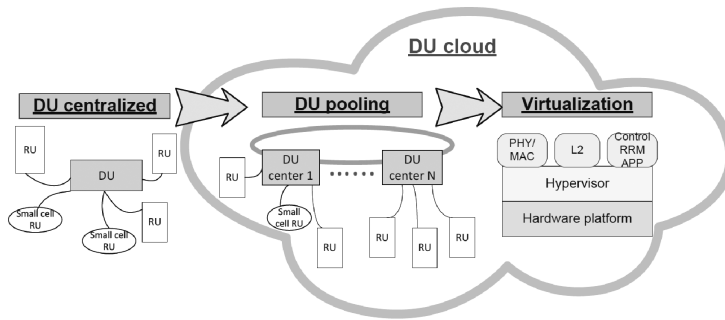


Figure 1.4: An exemplary C-RAN evolution.

Some of the high-level technical challenges for C-RAN are as follows [12]. Wireless signal processing requires strict real-time constraint in the processing. Baseband radio processing on a general purpose processor might be virtualized by Soft Defined Radio (SDR) techniques. Within a physical eNB virtualizing multiple logical RAN nodes from different mobile network systems, the processing resources must be dynamically allocated to higher-load logical RAN node keeping real-time scheduling and strict processing delay and jitter. DU pool must have a high-bandwidth and low-latency switching function with necessary data formats and protocols to interconnect among multiple DUs. I/O virtualization or API between PHY layer accelerator and standard IT platform must be addressed to access. Especially for C-RAN, higher consolidation of RRHs to a DU pool with higher I/O can benefit from a higher statistical multiplexing effect.

C-RAN can be evolved in step-by-step stages [22]. Figure 1.4 illustrates an exemplary evolution of C-RAN. In the figure, the first step is DU centralization, where DU can be in one location, and RF sites are connected to the DU using high-speed low-latency links. The second step is DU pooling, where multiple DUs are pooled and resources are not dimensioned by peak of individual DU site, but aggregated by the pool. The third step is virtualization of RAN, where the processing resources are virtualized and application independent of the hardware. The virtualization can use a hypervisor on top of the hardware platform, and different layers of the functions such as PHY/MAC, layer 2 (L2), radio resource management (RRM), and applications (APP) can be performed virtually, independent of the hardware.

1.3.3 Adaptive and Self-Organized RAN with Drop-and-Play Small Cells

Today's small cell, such as HeNB, is in a category of plug-and-play. This can reduce the cost of a planned network, where a lot of field labor, manual configuration, etc., can be minimized. With the development of SON, more and more small cells, not limited to HeNB, but also pico cells and so on, can be plug-and-play. SON can sup-

port self-configuration, self-optimization, and self-healing. For example, a cell can have self-configuration on the physical cell identifier, rather than getting a planned one as in a planned network. A cell can configure many parameters on its own, unlike in the traditional planned network. A cell can also have an automatic ON and OFF switch based on the current load. For example, if everyone is in the office campus, the HeNB can be off, while the small cells in office campus can be on, and if everyone is home, the office campus small cells can be off, while the HeNBs can be on.

The plug-and-play idea fits very well in terms of traffic load distribution, where the plug action can be related to the traffic demand, or the cluster-based UE distributions. Beyond plug-and-play, the future of small cells is place-and-play, or drop-and-play, where no wire is needed for the small cells, with the advances in wireless backhaul and energy harvesting. The SON functions should be enhanced to support the future drop-and-play deployment.

High-speed wireless backhaul is rapidly becoming a reality for small cells, which eliminates the need for wired connections. In another advance, the possibility of having a self-powered eNB is becoming realistic due to several parallel trends. First, eNBs are being deployed evermore densely and opportunistically to meet the increasing capacity demand. Small cells cover much smaller areas, and hence require significantly smaller transmit powers compared to the conventional macrocells. Second, due to the increasingly bursty nature of traffic, the loads on the eNBs will experience massive variation in space and time. In dense deployments, this means that many eNBs can, in principle, be turned off most of the time and only be requested to wake up intermittently based on the traffic demand. Third, energy harvesting techniques, such as solar power, are becoming cost-effective compared to the conventional sources. This is partly due to the technological improvements and partly due to the market forces, such as increasing taxes on conventional power sources, and subsidies and regulatory pressure for greener techniques. Therefore, being able to avoid the constraint of requiring a wired power connection or a wired backhaul is even more attractive, since it would open up entire new categories of low-cost place-and-play, or drop-and-play deployments, especially of small cells [14].

Figure 1.5 illustrates an exemplary architecture for drop-and-play deployment. In the figure, transient eNB (TeNB) is an access point or eNB without wire, which uses wireless backhaul to be connected to the core network, and is self-sufficient on its energy use via energy harvester. TeNB can apply drop-and-play. TeNBs can be deployed to increase the deployment density of the wireless network. TeNB can only be turned on for a small portion of the time. In other words, the duty cycle of the TeNB is low. Due to the low duty cycle of the TeNBs, preferably only a small number of the TeNBs in the network are turned on at a time. When a TeNB is turned on, it establishes a wireless backhaul link to the core network via hubs or other eNBs such as macro/micro/pico eNBs. A TeNB can also establish multiple links with multiple eNBs or hubs. Once the backhaul link is established, a TeNB can then provide an access link to UEs. The presence of TeNBs increases the deployment density of the network and thus can increase the capacity and coverage of the access link.

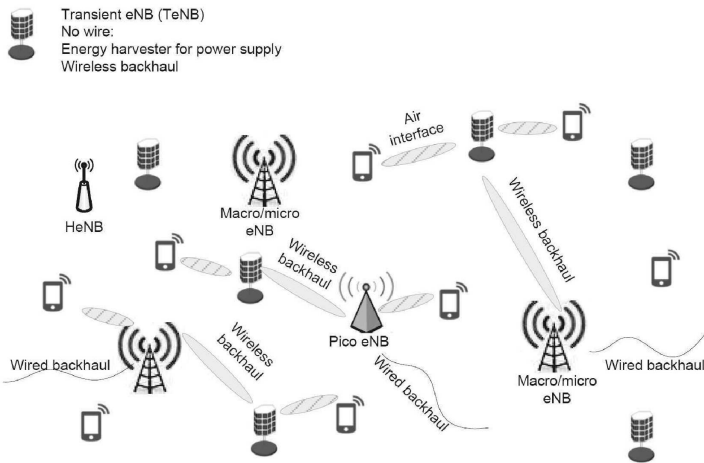


Figure 1.5: An exemplary architecture for drop-and-play deployment.

The network architecture as described above increases the robustness of the backhaul network. For example, if the communication link of one path of the wireless backhaul is congested or disrupted, TeNB can adapt the beamforming of its antenna array to establish communication via another path of the wireless backhaul. A TeNB can also have concurrent multiple paths for the wireless backhaul communication, or it can maintain multiple paths at the same time while communicating on one path at a time. SON functions can be enhanced to support route selection, establishment, and reroute.

In a TeNB, there is an energy generation module, an energy storage module (e.g., a battery), and a communication module, among others. The energy generation module can be either a solar power module, a wind power module, or power generation modules using other energy harvesting techniques. The power generated by the energy generation module can be fed either directly to the communication module or to charge the battery. The battery can then in turn power the communication module. The low duty cycle of the TeNB allows the energy generation module to be sufficiently small to ensure a small form factor of the overall device.

TeNB can also be an access point that can be turned on for a flexible portion of the time. The ON time of TeNB can be large or small. The duty cycle can be flexible. The duty cycle can be configured, indicated, updated, and sent to the other network entities, such as eNBs, UEs, backhaul hubs, etc. The network can configure or update the duty cycle based on considerations in the network such as load, distribution of the UEs, etc. This needs enhanced SON support.

The battery level of TeNB, the charging speed, etc., can be indicated and sent to the other network entities. The battery level of TeNB, the charging speed, etc., can be used as one of the factors to decide the route of the wireless backhaul, or for the UE to decide whether to access the TeNB. For example, when the battery level of

a TeNB is low, a UE may not choose to connect to the TeNB; rather, the UE may choose to connect to another TeNB nearby with longer battery life. Enhanced SON functions can be used to support such.

The TeNB can follow certain algorithm and triggering conditions to turn on or turn off. The ON/OFF switch can be dependent on the battery level, the charging speed, the traffic load, the traffic distribution, the price of the energy of the power grid, and so on. As many parameters can affect the ON/OFF switch, SON functions can be enhanced to support ON/OFF switches. Different modules in a TeNB can turn on or turn off at different times. The communication module of a TeNB can be turned on (or become active in serving UEs) via a variety of mechanisms. Note that the energy generation module of a TeNB can work when a TeNB becomes idle or active.

1.4 Future Evolved Core Network

In this section, future architecture for EPC is discussed, including Mobile SDN and network virtualization in EPC.

1.4.1 Mobile SDN

Mobile SDN (referred to as MobiSDN) is a type of networking for mobile networks where the control plane of the network is physically separate from the forwarding plane (or the data plane), and the data plane uses hardware including radio hardware (such as eNBs), servers, and switches. Network intelligence is (logically) centralized in software-based controllers, which maintain a global view of the network.

MobiSDN support in the EPC has gained a lot of attention from operators and vendors. The core network may have a MobiSDN architecture in the future [17, 28]. In addition, the interface between eNB and the core network may also be impacted, introducing the need for MobiSDN-capable eNB. MobiSDN allows further development of smart edge solutions, such as content caching and local APP server hosting [28]. With MobiSDN, more and more use cases for additional revenue or value added services can be provided.

MobiSDN is comprised of smart edge and cloud EPC [28]. Figure 1.6 illustrates an example of the architecture of MobiSDN. The details of Figure 1.6 are explained as follows.

Smart edge: The smart edge includes SDN-capable eNBs. The edge controller is also SDN based, whose function can be part of the central controller. The edge server may be co-located with the eNB. The smart edge has three main functionalities: distributed computing, distributed file system, and networking controller. Distributed computing enables different processing capabilities at the edge, including computation load balancing and programming transparency. The distributed file system can support distributed storage, cache sharing, content search, etc. The network controller is based on SDN with programmable routers, flexible policy checking, and is friendly

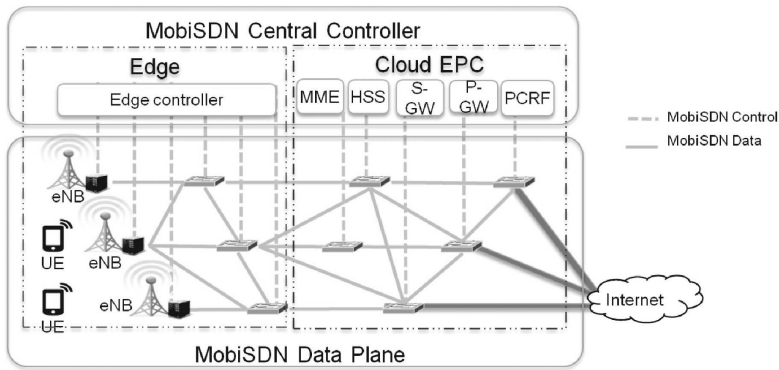


Figure 1.6: An exemplary architecture for MobiSDN.

to middle boxes. These functions are inevitable and needed, considering the volumes of mobile data and huge mobile video traffic demand.

Cloud EPC: The cloud EPC is SDN capable. It uses SDN switches and servers as the hardware. The data plane is based on SDN switches that provide data forwarding. The control plane is based on the MobiSDN central controller. The controller can take on functions including those provided by the MME, HSS, S-GW, P-GW, PCRF, etc. These functions can be applied within each of the SDN switches. The switches are also involved in routing and running transport protocols.

Although here the functions in the data and control planes are described by using the names of MME, HSS, S-GW, P-GW, PCRF, etc., there may or may not be these network entities anymore. For example, the MME, PCRF, and HSS can be absorbed in the MobiSDN central controller, while some of the functions of S-GW and P-GW will be in the data plane (e.g., the MobiSDN switches) and some will be in the control plane (absorbed in the MobiSDN central controller).

There are two essential differences with MobiSDN compared to the state-of-the-art architecture. First, MobiSDN has a clear separation of data plane and control plane. Second, MobiSDN is very flat, contrasting to the existing hierarchical architecture wherein the P-GW can be the bottleneck as the node connecting to the Internet. MobiSDN has the following advantages: more flexible routing and flows, flexible middle boxes, keeping the content in the edge, offloading traffic from the core network, handover with less overhead, flexible radio resource management and scheduling, etc.

1.4.2 Network Virtualization in EPC

Network virtualization in EPC is a technology by which resources of the network (for example, the hardware) can be virtualized and used transparently. It enables the creation of a competitive environment for the supply of innovative third-party network

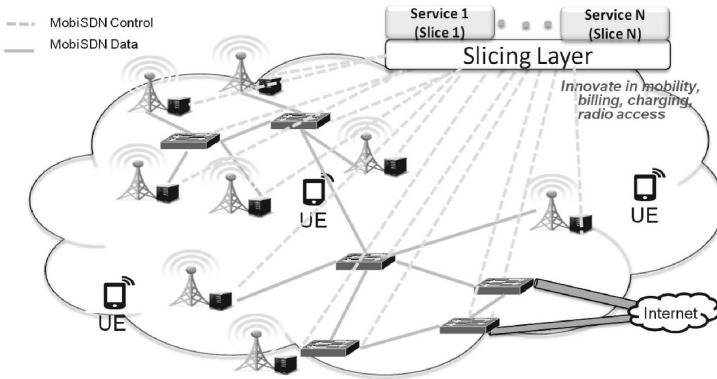


Figure 1.7: An exemplary virtualization in MobiSDN.

applications by unlocking the proprietary boundaries of mobile base station nodes. Virtualization is a good tool to achieve the paradigm of soft networking. Network function virtualization is under discussion [25]. The virtualization of eNBs in Section 1.3.2 is an example of virtualization in the RAN, and the current section focuses more on the virtualization in the EPC.

One example to achieve network virtualization in EPC can be by slicing the flow space, such as by using a hypervisor. All the network hardware can be used as shared infrastructure by all the slices. A slice may consist of part or all of different infrastructure (e.g., eNBs, switches) elements. Each slice can be used to provide a different value-added service. Since each slice can be flexibly and independently managed, it allows for the easy introduction of new revenue opportunities without additional hardware complexity costs [17].

Figure 1.7 illustrates an example of virtualization in MobiSDN. A hypervisor can be used, to support the slicing layer, on top of which different services can be supported. Each slice can be used by a different service and virtually correspond to part or all of the hardware at the physical layer.

Some examples of using virtualization for value-added services are provided as follows. One example is efficient real-time communication for enterprises with multiple-campus support. An enterprise that has multiple campuses in different geographical areas may be interested in deploying low-latency interactive wireless services, such as allowing employees to have high QoS interactive wireless video conferencing, doc collaboration query processing, etc., on top of the existing wireless network. The operator can provide this service by using MobiSDN and a slice consisting of the eNBs close to the campuses (not all the base stations are needed, which simplifies the networking) and switches involved. An advantage of this approach is that it creates new services and more revenue for operators while also reducing management complexity.

Another example is for supporting stadiums and other large venues. A slice can consist of local eNBs within the stadium and switches involved, to provide local content and services. Users can pay for premium service with better QoS since congestion is often an issue in these scenarios, or participate in certain events such as video contests, etc. Content, such as players' introduction, video replay of exciting game highlights, etc., can be stored in the local cache server. The UE can also upload its captured video or other content to the local server, to share with other local users. The venue or operator may incentivize content uploading and sharing by offering rewards for users who provide very high quality video clips that become popular.

1.5 Conclusion

This chapter provides discussions on the future cellular network architecture. For the future RAN, potential technologies, including the support for a UE to connect concurrently to multiple eNBs and associate diverse traffic to the eNBs, Cloud-RAN, and advanced SON to support drop-and-play small cells, are discussed. For the future EPC, potential technologies, including mobile SDN and network virtualization, are discussed. All these technologies provide great potential to improve today's cellular network architecture and leverage the limitations of today's architecture.

The future cellular network architecture may not be limited to what is discussed in this chapter. For example, the architecture of interworking of cellular and Wi-Fi can be further improved, as currently discussed in [9]. For another example, the architecture of supporting device-to-device communications can be consolidated to today's cellular architecture, as currently discussed in [6].

All in all, the future cellular network architecture will be supporting more use cases and be more adaptive, more optimized, more efficient, more cost effective, and easier for network management than today's architecture.

References

1. 3GPP TR 36.842 v0.2.0. Study on small cell enhancements for E-UTRA and E-UTRAN—Higher-layer aspects. 3GPP TSG RAN, 2013.
2. 3GPP R1-083813. Range expansion for efficient support of heterogeneous networks. 3GPP, TSG-RAN Work Group 1 (WG1) 54bis, Qualcomm Europe, 2008.
3. 3GPP R1-101505. Extending Rel-8/9 ICIC into Rel-10. 3GPP, TSG-RAN WG1 60, Qualcomm, February 2010.
4. 3GPP RP-130888. LTE TDD-FDD joint operation. 3GPP TSG RAN, 2013.
5. 3GPP TR 36.819 v11.1.0 (2011-2012). Coordinated multi-point operation for LTE physical layer aspects (release 11). 3GPP TSG RAN, December 2011.

6. 3GPP TR 36.843. Feasibility study on LTE device to device proximity services—radio aspects. 3GPP TSG RAN, 2013.
7. 3GPP TR 36.872. Small cell enhancements for E-UTRA and E-UTRAN—Physical layer aspects. 3GPP TSG RAN, 2013.
8. 3GPP TR 36.932 v12.0.0. Scenarios and Requirements for small cell enhancements for E-UTRA and E-UTRAN (release 12). 3GPP TSG RAN, December 2012.
9. 3GPP TR 37.834. Study on WLAN/3GPP radio interworking. 3GPP TSG RAN, 2013.
10. 3GPP TR 36.847. LTE time division duplex (TDD)—Frequency division duplex (FDD) joint operation including carrier aggregation (CA). 3GPP TSG RAN, 2013.
11. China Mobile. C-RAN: Strategy, trial and future considerations. Proceedings of IWPC, December 2012.
12. GS NFV 009 v015. Network function virtualization: Use cases. ETSI, June 2013.
13. H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam. Distributed α -optimal user association and cell load balancing in wireless networks. *ACM Transactions on Networking*, 20(1):177–190, 2012.
14. H. S. Dhillon, Y. Li, P. Nuggehalli, Z. Pi, and J. G. Andrews. Fundamentals of base station availability in cellular networks with energy harvesting. Proceedings of the IEEE Globecom, December 2013.
15. ITU, Report M.2135. Guidelines for evaluation of radio interface technologies for IMT-Advanced. 2008.
16. J. Kempf, B. Johansson, S. Pettersson, and H. Luning. Moving the Mobile Evolved Packet Core to the cloud. Fifth International Workshop on Selected Topics in Mobile and Wireless Computing, 2012.
17. L. E. Li, M. Mao, and J. Rexford. Towards software-defined cellular networks. First European Workshop on Software Defined Networking, October 2012.
18. L. Liu, Y. Li, B. Ng, and Z. Pi. Radio resource and interference management for heterogeneous networks. In *Heterogeneous Cellular Networks*, John Wiley & Sons, 2013.
19. Open Networking Foundation. www.opennetworking.org.
20. Open Networking Summit. www.opennetsummit.org.
21. OpenDaylight. www.opendaylight.org.

22. Project Proposal. C-RAN: Centralized processing, collaborative radio, real-time cloud computing clean RAN system. NGMN Alliance, March 2011.
23. Telstra and Ericsson. Service provider SDN meets operator challenges. Open Networking Summit, April 2013.
24. TS 36.300 v11.6.0, 3rd Generation Partnership Project (3GPP); Technical Specification Group Radio Access Network (TSG RAN). Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); overall description; stage 2. June 2013.
25. White paper. Network functions virtualization—Introductory white paper. ETSI, December 2012.
26. White paper. Cisco visual networking index: Forecast and methodology 2012–2017. May 2013.
27. Y. Li, A. Maeder, L. Fan, A. Nigam, and J. Chou. Overview of femtocell support in advanced WiMAX systems. *IEEE Communications Magazine*, July 2011.
28. Y. Li, M. Dong, D. Choe, T. Novlan, C. Zhang, and G. Morrow. MobiSDN: Vision for Mobile Software Defined Networking for Future Cellular Networks. In Proceedings of Globecom 2014, Industry Forum.
29. Y. Li, Z. Pi, and L. Liu. Distributed heterogeneous traffic delivery over heterogeneous wireless networks. Proceedings of the IEEE ICC, June 2012.