# Chapter 11
# Rethink Ring and Young: Green and Soft RAN for 5G

**Chih-Lin I, Jinri Huang, Ran Duan, Gang Li and Chunfeng Cui**

**Abstract** This chapter discusses one of the key design principles for 5G systems: “No More *Cells*” (NMC) [1]. NMC transfers the traditional cell-centric network design to a user-centric design principle. It is pointed out that NMC realization could be facilitated by the Cloud RAN (C-RAN) architecture which pools the processing resources and virtualizes “soft” BBUs and various applications on demand. The major challenges for C-RAN, including the transport networks to connect the resource pool and the remote sites as well as virtualization with potential solutions, are analyzed in detail. Various fronthaul solutions, including Common Public Radio Interface (CPRI) compression, single-fiber bi-direction, as well as wavelength division multiplexing (WDM) technology, are demonstrated and verified through our extensive field trials. In addition, the feasibility of general purpose processor (GPP) platform adoption in baseband processing with optimized virtualization implementation is functionally demonstrated and initially verified in terms of interruption time through prototype development implemented with a commercial LTE protocol stack.

## 11.1 Introduction

The concept of cellular systems was proposed in 1947 by two researchers from Bell Labs, Douglas H. Ring and W. Rae Young. Since the first generation of cellular standards, this cell-centric design has been maintained through every new generation of standards including 4G. The nature of a homogeneous cell-centric design is that cell planning and optimization, mobility handling, resource management, signaling and control, coverage, and signal processing are all assumed to be done either for or by each base station (BS) uniformly.

C.-L. I (✉) · J. Huang · R. Duan · G. Li · C. Cui
Green Communication Technology Research Center,
China Mobile Research Institute, Beijing, China
e-mail: icl@chinamobile.com

© Springer International Publishing Switzerland 2017

M. Tornatore et al. (eds.), *Fiber-Wireless Convergence in Next-Generation Communication Networks*, Optical Networks,
DOI 10.1007/978-3-319-42822-2_11

To meet the rapid traffic growth (potentially 1000× by 2020), network densification is viewed as a major way to achieve the targeted increase in throughput. Further, heterogeneous network (HetNet) deployment is widely adopted. Thus, from the network's perspective, diverse types of BSs with different coverage, transmit power, frequency bands, etc., are introduced, such as Macro, Micro, Pico, and Femto. Also, from the users' perspective, traffic fluctuation is more significant than before, taking into account the emerging millions of mobile data applications. Therefore, in practical deployment, it is clear that the current system design for homogeneous networks is not compatible with traffic variations and diverse radio environments. Conventionally, radio resources are allocated semi-statically from the standpoint of the network rather than the user equipment (UE). This approach lowers resource utilization and wastes power. Poor performance at the cell edge also severely influences the consistency of the user experience. In addition, user mobility causes handover to occur frequently, especially under dense small cell deployment [2]. Consequently, radio resources are re-assigned with complex neighboring cell monitoring algorithms and a high signaling overhead. Moreover, frequent handover failures degrade the user experience significantly.

Relays, distributed antenna systems (DASs), coordinated multipoint processing (CoMP), and HetNets have been implemented as short-term solutions for these issues. While relays and DASs are mainly used for coverage extension, HetNets are deployed mainly for capacity improvement. CoMP has been intensively investigated by academia, industries, and standard bodies like 3GPP and WiMax, in which inter-BS joint processing and/or coordination is sought for enhanced cell-edge and cell-average performance. Note that in the above initial efforts, the cell-centric network operation was hardly changed.

## 11.2 No More Cells: One Key 5G Vision

Recently, Beyond Cellular Green Generation (BCG2) [3], liquid cells, soft cells, and phantom cells [4] have surfaced as potential radio access architectures. These paradigms all lead to the principle of "No More *Cells*" [1]. Different types of information can come from different sites, and the set of sites used for transmission is transparent to the UE.

5G design of the user-centric radio network should start with the principle of "No More *Cells*," departing from cell-based coverage, resource management, and signal processing. User demand, rather than cell, should be centric to network radio resource assignment and processing. All nearby radio access points with diverse frequency bands, transmit power, and coverage could serve one user. Moreover, the available radio resources from multiple access points could be dynamically scheduled for coordinated multipoint transmission, and the selection of control/user plane and uplink (UL)/downlink (DL) channels, respectively.

The concept of "No More *Cells*" (NMC) is user-centric with amorphous cells, decoupled signaling and data, and decoupled DL and UL. For example, a macro-BS with lower frequency and wider coverage would become a signaling BS, while small cells with higher frequency and overlapped coverage would be data-only BSs [5, 6]. In a HetNet scenario, the small cell is within the coverage of a macrocell. Even if the small cell has no traffic, it cannot be turned off in the traditional cell paradigm. But with a control and data decoupling scheme, the macrocell is responsible for control and the small cell only for data. Thus, when there is no data traffic in the small cell, it can be completely turned off to save energy. New users can access the macrocell, and then the macrocell can coordinate with the small cell for possible data transmission. With signaling and data decoupled, the mobility robustness can be improved since handover signaling overhead is reduced with a more stable signaling connection when employing a macrosignaling BS. Additionally, spectrum usage in small cells will be significantly enhanced, due to the relaxed need for transmission of control information and reference signals from small cells.

NMC is also user-centric with decoupled UL and DL. This decoupling is deemed to facilitate better resource allocation between cells. This can be illustrated in the following example: Consider two cells where cell 1 is heavily loaded in the DL and cell 2 overloaded in the UL. In the traditional cell concept, if a UE is located at the cell boundary with symmetric data requirements, and the serving cell is cell 1, its DL requirement may not be satisfied. Conversely, if the UE device's serving cell is cell 2, its UL requirement may not be satisfied.

This situation of UL/DL asymmetry is becoming even worse under HetNet deployment since UL is always associated with DL based on the measurement of "best DL" reference signals. However, the nearby small cells with less reference signal power may be likely to provide a better UL connection. If there is a user-centric network design, the UE's DL can be from one radio access point and the UL from another access point, meeting the UE device's data requirement for symmetric transmission.

Moreover, CoMP with coordinated scheduling and multipoint transmission can be utilized to improve the split of the control/user planes and UL/DL channels. For example, under dense HetNet deployment, UE may anchor to one macro/signaling BS to establish a stable signaling connection and attach to small cells for data connections. Furthermore, UE may select UL and DL data connections separately from different radio access points and the DL data connection may be served dynamically by multiple radio access points.

With the introduction of ultra dense network (UDN) in 5G, the radio environment is becoming more complicated than ever due to extensive overlapped coverage. Thus, more radio channel information between radio access points nearby should be shared in real-time and more joint cooperation between neighboring access points is required to implement control/user plane and UL/DL channel selection from one or multiple points. This is not feasible in a traditional radio access network, because too much inter-BS information sharing is incurred, including dynamic user channel state information and scheduling information. Fortunately, with the emergence of C-RAN [7, 8] as shown in Fig. 11.1, many technologies necessary for the realization of the NMC concept can be facilitated. In addition, system-level and even multi-radio access technology (multi-RAT) optimization can also be made possible.

In the following sections, the concept of C-RAN will be elaborated with its key features and advantages. The major challenges and potential solutions will also be discussed. Finally, we will present recent progress in this area in terms of extensive C-RAN field trials and prototype development.

## 11.3 Cloud RAN: The Key Enablers to NMC

### *11.3.1 The Concept of C-RAN*

Figure 11.1 shows the basic concept of C-RAN, while Fig. 11.2 illustrates more architecture details. A C-RAN system centralizes different processing resources to form a pool in which the resources could be managed and dynamically allocated to different applications on demand. The key design principle of C-RAN is to support various kinds of applications running on the same hardware platform. The key enabler for this is the virtualization technology widely used in modern data centers. With virtualization, standard IT servers are used as the general platform with computation and storage as the common resources. As shown in Fig. 11.2, a C-RAN system runs different applications on top of the servers in the form of virtual machines (VMs). The indispensable applications in C-RAN are those that can realize different radio access technologies (RATs) including 2G, 3G, 4G, and,

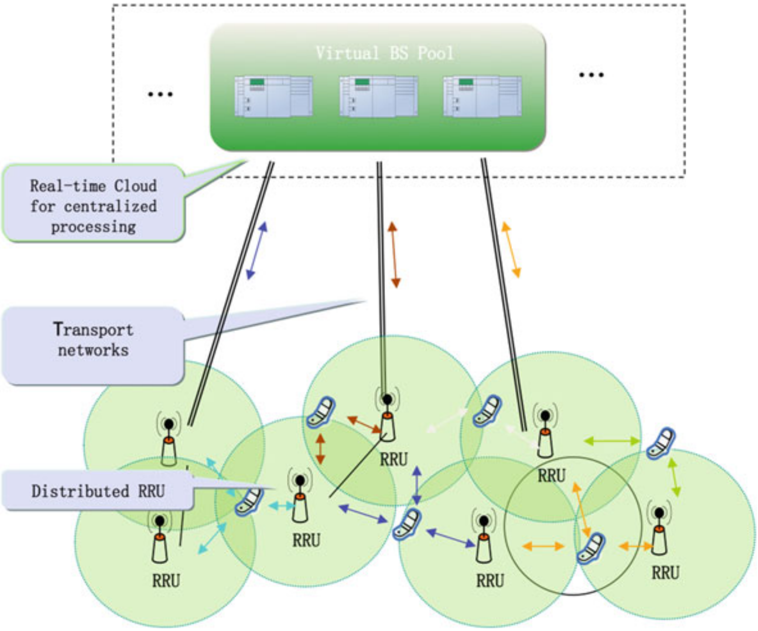


**Fig. 11.1** Illustrative C-RAN concept

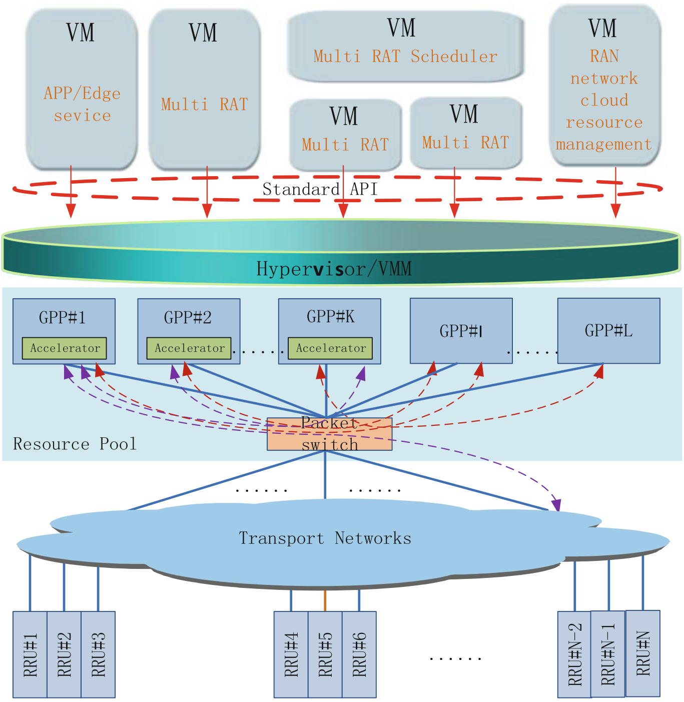


**Fig. 11.2** C-RAN architecture

eventually, 5G. Additional user applications such as content delivery networks (CDN) and web caches can also be deployed on the open virtualized platform. In addition, the C-RAN platform provides a set of standard application program interfaces (APIs), which open the opportunity for new service provisioning and deployment. In this way, C-RAN is no longer a single RAT processing entity but rather a platform for the co-existence of diverse services.

The C-RAN architecture consists mainly of three parts.

- **Resource pool**: In C-RAN, all the resources are pooled and virtualized so that they can be delegated for use in different applications. One indispensible application in C-RAN is “soft” BBU. A soft BBU is a BBU instance in a traditional network where processing resources and capabilities are dynamically allocated and reconfigured based on real-time conditions (e.g., traffic status). In addition, a prerequisite for a pool in C-RAN is an inter-connection switching network of high bandwidth and low latency. This switching mechanism realizes an inter-connection of different computation nodes and enables efficient information exchange among them.
- **Remote Radio Units** (**RRU**)**/antenna sites**: Design of RRU or antenna networks is relatively independent of the central pool. The remote RRU networks could be the same as those in traditional systems for easiest migration.

Alternatively, potential 5G antenna or RRU technologies such as large-scale antenna systems (LSASs) could also be supported in C-RAN. In that case, there may be an impact on the interface between the central pool and the remote sites.

- **Transport networks**: A transport network provides a connection between the resource pool and the remote sites. It could be of different forms depending on practical situations and scenarios. Some examples include direct fiber connection via dark fiber, microwave transmission, and fiber transport networks.

Traditional wireless systems can easily migrate into a C-RAN architecture in two stages. The first stage realizes centralization by aggregating existing BBUs from different locations into one equipment room. In the second stage, virtualized general purpose platforms (GPPs) should be introduced to replace vendor-specific platforms.

### *11.3.2 C-RAN Features*

Although C-RAN might at first appear to be nothing more than the centralization of BBU, centralization is just the first step toward a complete C-RAN that includes several other features:

- **Advanced Technology Facilitation**: The high-bandwidth and low-latency inter-connection switching network facilitates the efficient and real-time information exchange among different computation nodes. As a result, many technologies that are difficult to implement in traditional architectures, especially joint processing and cooperative radio, can benefit from this feature and thus will become viable in a C-RAN context.
- **Resource Virtualization/Cloudification**: Unlike traditional RAN systems in which computation resources are limited within one BBU and therefore cannot be shared with other nodes, in C-RAN these resources are aggregated on a pool level and can be flexibly allocated on demand. This feature, very similar to the cloud and virtualization concepts in data center, is called resource "cloudification."
- **"Soft" BBU**: Traditional wireless equipment is developed based on proprietary platforms and possesses only "hard" fixed capabilities designed for carrying peak traffic. Contrary, in a C-RAN BBU pool, through resource cloudification, a BBU is of a soft form, which means that the capability of a soft BBU could be dynamically reconfigured and adjusted. In this way, resource utilization efficiency can be increased considerably.
- **Generalization of platform**: Generalization of platform is another essential feature of C-RAN. The use of GPPs not only reduces the procurement cost for operators but more importantly lays down the basis for the implementation of virtualization technology.
- **Openness**: C-RAN is also "open" in the sense that C-RAN is designed to provide a set of standard APIs to outside parties to encourage new service development on the edge and to provide interoperability.

### *11.3.3 Advantages of C-RAN*

Several advantages can be directly derived from the C-RAN architecture:

- **Total cost of ownership (TCO) reduction**: TCO reduction mainly comes from two sources. Firstly, centralization allows the aggregated computation nodes to share the same facilities, such as air-conditioning, which reduces the power consumption and therefore operating expenses (OPEX). Power consumption by air-conditioning usually accounts for over half of the total power consumption for operators. Secondly, centralization makes it easier to find a smaller number of central offices to accommodate the BBU node pool, which in turn can speed up network construction. Improved resource utilization efficiency due to resource cloudification also contributes to TCO reduction.
- **Improved system performance**: Internal high-bandwidth low-latency switching networks enable implementation of advanced joint processing techniques, which leads to system performance improvement.
- **Energy saving**: Similarly to TCO reduction, energy saving is attributed to two factors: centralization and resource cloudification. First, facility sharing in the same central office helps to reduce energy consumption. Second, reduction in equipment usage due to improved resource utilization efficiency by cloudification reduces the total power consumption.
- **Improved resource utility efficiency**: This benefit is mainly from resource cloudification.
- **Facilitation of service deployment on the edge**: A C-RAN network covers a larger area and serves more users than a traditional single BS. Making use of this, it is possible to move services to or directly deploy new services on the RAN side. In this way, user experience could be improved and backhaul pressure could be relieved.

## 11.4 Challenges and Potential Solutions for C-RAN Realization

### *11.4.1 Challenges on Transport Networks for Centralization*

Centralization is the critical first step required in order to realize all the features of C-RAN. Centralization aggregates different BBUs which traditionally are located in geographically disparate places into one central office with shared facilities.

Centralization of certain carriers may consume a significant number of fiber resources if using a dark fiber solution, i.e., direct fiber connection. For example, in a Time Division Long-Term Evolution (TD-LTE) system with 20 MHz bandwidth and RRUs equipped with 8 antennas (most common scenario in the China Mobile (CMCC)'s network), the CPRI [9] data rate between one BBU and one RRU for one TD-LTE carrier transmission is as high as 9.8 Gbps. When considering both UL and DL, then 4 fiber connections would be required with 6-Gbps optical modules. Since usually one site consists of three sectors with each supporting at least one carrier, the number of fiber connections for one site is as high as 12. When the centralization scale becomes larger, the amount of fiber needed will be greatly increased, resulting in an implementation which is difficult to achieve for most operators due to limited fiber resources. As a result, a dark fiber solution is usually not recommended. Instead, an efficient transport solution is needed to address the fiber consumption issue in order to achieve C-RAN large-scale deployment.

When a transport network is adopted, there will be a new challenge resulting from the dynamic BBU-RRU mapping requirement. In C-RAN, the BBU is soft and can be dynamically changed. The relationship between the BBU and RRU is no longer a fixed one-to-one correspondence as in traditional RAN systems. Therefore, the transport network should be able to support dynamic mapping and routing for data between the pool and the remote RRU. Software-defined networks (SDNs) could be one of the promising technologies to provide this capability. The basis of SDN is to separate the control plane from the forwarding plane and change the forwarding behavior through programming on the control plane [10]. In this way, SDN greatly improves the transport efficiency and flexibility.

### *11.4.2 Potential Fronthaul Solutions*

In the subsequent subsections, the transmission between the BBU and the RRU in C-RAN is defined as fronthaul transmission (compared with traditional backhaul transmission between the BBU and the core network).

The fronthaul transmission technology is of decisive significance to C-RAN large-scale deployment. As more operators are paying attention to C-RAN, more resources are being committed to the issue. There have been some notable breakthroughs recently:

- **CPRI compression**. With the maturity of CPRI compression, several vendors have commercially realized 2:1 compression with lossless performance, potentially reducing fiber consumption by 50 %. In addition, single-fiber bi-direction (SFBD) technology allows simultaneous UL/DL transmission on a single fiber, which further halves fiber consumption. Combining CPRI compression and SFBD can reduce fiber consumption threefold. CMCC has successfully verified both technologies in C-RAN TD-LTE field trials. More information can be found in Sect. 11.5.1.
- **WDM solutions**. Since WDM technology is sufficiently mature, vendors can develop WDM equipment tailored to fronthaul transmission within a short period of time. Currently, a few operators have adopted this solution to enable the large-scale C-RAN deployment. Some commercial products can support as many as 60 2.5-Gbps CPRI links in one pair of fibers. 1+1 or 1:1 ring protection

is also supported and several low-data-rate links can be multiplexed into one high-data-rate link. The main issue with the solution lies in its high cost, which hinders its large-scale deployment.

- **OTN** (**optical transport network**) **solutions**. Compared with a WDM solution, OTN provides more powerful O&M capability, longer reach and flexible routing. In addition, the open interface and standard protocols of OTN help to lower the cost and decrease development time. Some vendors have suggested integrating OTN functions into optical modules rather than using active line cards, which can simplify network deployment and maintenance to a large extent.
- **Millimeter microwave transmission**. In some scenarios, it is too expensive, or even impossible to deploy fiber. In that case, microwave transmission may come to play a role as the last-100-m fronthaul solution. 60 GHz is currently the most common frequency band for millimeter microwave and can be implemented under loose regulation in many countries. 60 GHz offers wide bandwidths, and thus it is easy to get channels with 250 MHz or greater bandwidth. With simple modulation techniques, it is easy to achieve over 1-Gbps transmission rate within the 100–400 m range. For LTE RRU with 20 MHz bandwidth and 2 antennas, the data rate after 2:1 compression is less than 1 Gbps and can be transmitted via millimeter microwave. 5 GHz millimeter microwave products have recently entered the market and can support the fronthaul transmission of 20 MHz LTE with 8 antennas.
- **CPRI redefinition**. The basic idea behind CPRI redefinition is to move a partial set of physical-layer functions to the RRU side in order to reduce the required data rate between the BBU and the RRU. There can be several possibilities for the partition of functions. By carefully designing the partition scheme, the data rate between the BBU and the RRU can become elastic and vary with real user traffic, as opposed to the traditional case in which the I/Q stream is constant even when there is no real traffic. This feature helps not only to reduce the capacity requirement on the switching network within the BBU pool but also to reduce the switching latency. In addition, the data can now be encapsulated in the form of packets rather than a constant stream and therefore can be transmitted by a packet switching protocol, such as Ethernet which enjoys the benefits of improved flexibility and improved switching efficiency. In Sect. 11.6.2, this idea is expanded.
- Wavelength Division Multiplexing-Passive Optical Network (**WDM-PON**). Using WDM-PON as a C-RAN, fronthaul has recently been discussed as an alternative in the Full Service Access Network (FSAN) and ITU-T Q2 working group. The basic idea is to make use of the rich fiber resource deployed for Fiber-to-the-x (FTTx) and design a new technology based on the combination of the low-cost PON and WDM for CPRI transmission. WDM solutions adopt colored optical modules, which raise the bar in installation of optical modules, maintenance, and storage. In comparison, WDM-PON targets use of colorless optical modules, which helps to greatly simplify installation, maintenance, and storage. In addition, WDM-PON claims such advantages as cost reduction,

saving on fiber consumption, and flexible topology support. Despite being at its initial stage, WDM-PON can become one of the most efficient fronthaul solutions for C-RAN in the long run.

### *11.4.3 Challenges on Virtualization Implementation to Realize Resource Cloudification*

C-RAN's core feature is resource cloudification in which processing resources can be dynamically allocated to form various functional nodes (e.g., BBU). In addition, with C-RAN, all the diverse applications and services should operate independently and simultaneously on the same platform in the form of software. To enable this, it is widely believed that virtualization is one of the key components. In fact, this design philosophy is exactly the same as that employed in network function virtualization (NFV). Proposed by China Mobile and 12 other operators in a joint white paper under ETSI NFV ISG [11], NFV has been widely accepted in industry as one of the most important technologies in order to realize 5G. Incorporating C-RAN as an essential part, the basic idea of NFV is to "consolidate many network equipment types onto industry-standard high-volume servers, switches and storage, which could be located in data centers, network nodes and in the end-user premises" [11]. Prior to the establishment of the NFV ISG (Industry Specification Groups) under ETSI (European Telecommunications Standards Institute), its original defined scope focused on core networks. We propose instead that it should be extended to include RAN, which would make the NFV concept be truly end to end.

Virtualization, in its simplest form, is a mechanism to abstract hardware and system resources from a given operating system (OS). This is usually performed via hypervisor which separates an OS from the underlying hardware resources. Certain functionalities usually run in the form of VMs on top of the hypervisor. VMs are separated from each other and this separation provides independence and ensures security.

Despite the simplicity of the idea of using virtualization, the implementation is more difficult in practice. Wireless communication is distinct from IT data centers in that wireless communication has extremely strict requirements on real-time processing. For example, for LTE systems it is required that an acknowledgement/negative acknowledgement (ACK/NACK) must be produced and sent back to the UE/eNodeB (eNB) within 3 ms after a frame is received. Traditional data center virtualization technology with GPP cannot meet this requirement. Therefore, applying virtualization to BSs requires careful design and special optimization of key functional blocks.

To better describe the issues, let us take LTE C-RAN as an example. Then these challenges include, but are not limited to the following:

- **Meeting the real-time (RT) constraint for system performance**: The biggest challenge for RAN virtualization is to meet the strict RT constraint imposed by mobile signal processing. To partially address this, from the physical (PHY) layer perspective, a dedicated hardware accelerator is proposed to process computation-intensive function modules such as channel coding and decoding. However, when it comes to L2 and L3, this constraint still holds. Since L2 and L3 are mainly processed on GPP platforms with the software layers of hypervisors and OSs, the issue can then be translated to the optimization of the hypervisor and OS. The optimization of these components is not only to fulfill the RT requirement but also to minimize the overhead and guarantee system performance.
- **Meeting the RT requirement for VM management, especially for live migration**: In traditional live migration, the workload of one VM can live migrate to another. Although this idea still applies to C-RAN systems, it would become much more difficult when it comes to mobile communication implementations. The main issue lies with the high arrival rate of wireless frames, for example, one subframe per millisecond in LTE systems. This is often shorter than the migration interval from one VM to another. This will lead to a non-convergence of migration, resulting in failure. Moreover, when a source VM is switched off and the destination VM takes over, it usually requires an interruption of several hundred milliseconds, which, from a wireless communication point of view, means the inevitable loss of several hundred subframes. This is obviously intolerable for wireless systems.
- **I/O virtualization**: As mentioned above, C-RAN requires the use of an L1 accelerator to partially solve the RT problem. The data exchange between the accelerator and the upper-layer software could result in high I/O requirements. In a traditional virtualization environment, the data communication between the VM and the underlying hardware needs the hypervisor's intervention, which brings additional overhead and therefore results in further degradation of I/O performance, particularly the I/O throughput and the latency. Therefore, I/O virtualization techniques should be introduced to improve system I/O performance. Moreover, it is possible that VMs outnumber accelerators in C-RAN. In this case, a mechanism of I/O virtualization is needed to enable different VMs to share the accelerators without degraded performance.
- **Design on virtualization granularity**: There are several possible ways to compose a VM. Taking LTE L2/L3 virtualization as an example, there could be as many VMs as carriers, each VM dealing with one carrier. Alternatively, there could also only be two VMs with one dealing with all L2 functions and the other dealing with all L3 processing. Even a VM for one UE is possible. It is obvious that the VM granularity, or composition, can have different impacts on system performance and VM management and is thus worth careful study.
- **Evaluation of different hypervisor alternatives**: There exist in industry various hypervisor products including Xen, Vmware ESX, Oracle VirtualBox, and KVM [12]. Different hypervisors differ from each other in many aspects such as the supporting virtualization types, OS, architecture, core count, memory

capacity, and live migration. It would be of great value to evaluate which hypervisors are more suitable for RAN virtualization and how.

## 11.5 Recent Progress on C-RAN from China Mobile

In this section, we will introduce the recent progress on C-RAN activities at China Mobile. We have conducted extensive field trials, some of which being commercial LTE networks, to verify various fronthaul solutions. In addition, we also developed several sets of prototypes to evaluate the implementation of virtualization technology for wireless communication.

### *11.5.1 Field Trials on Centralization with Different FH Solutions*

The first step toward C-RAN is BBU centralization which is relatively easy to implement and can be tested with existing 2G, 3G, and 4G systems. In the past few years, extensive field trials have been carried out in more than 10 cities in China using commercial 2G, 3G and pre-commercial TD-LTE networks with different centralization scales. The main objective of C-RAN deployment in 2G and 3G is to demonstrate deployment benefits by centralization, including site construction speed-up and power consumption reduction.

In the city of Changchun, 506 2G-BSs in five counties were upgraded to a C-RAN-type architecture, centralized into several central sites. In the largest central site, 21 BSs were aggregated to support 101 RRUs with a total of 312 carriers. It was observed that power consumption was reduced by 41 % due to shared air-conditioning. In addition, system performance in terms of call drop rate as well as downlink data rate was increased using same frequency network (SFN) technology. More details on 2G and 3G C-RAN trials can be found in [7, 8].

While C-RAN trials in 2G and 3G demonstrated the viability and advantages of centralization, C-RAN trials in TD-LTE aimed to verify availability of technologies to reduce fiber consumption, including CPRI with 2:1 compression and SFBD which, as described in Sect. 11.4.2, allows simultaneous UL/DL transmission within the same fiber core.

When combining 2:1 CPRI compression and SFB together, fiber resources could be saved threefold with lossless performance.

To verify the two technologies in real networks, three field trials were carried out in three cities with similar configurations to each other as shown in Table 11.1. In these trials, commercial eNBs and Evolved Packet Core (EPC) were used together with test UEs.

**Table 11.1** System configuration of TD-LTE C-RAN field trial

| | |
|---|---|
| Frequency | 2.85 GHz |
| Bandwidth | 20 MHz |
| Frame structure | UL/DL configuration type 1<br>• Normal CP<br>• Special subframe configuration type 7 (DwPTS:GP:UpPTS = 10:2:2)<br>• DwPTS for data transmission |
| CPRI | 2:1 compression |
| Optic module | Single-fiber bi-direction |
| UL | SIMO |
| DL | Adaptive MIMO |
| QCI | 9 |
| Scheduler | PF |

**Table 11.2** Throughput comparison b/w with and without compression plus SFBD

| | RSRP* | SINR** (dB) | DL (Mbps) | | UL (Mbps) | |
|---|---|---|---|---|---|---|
| | | | w/ | w/o | w/ | w/o |
| Near point | (−75, −85) | >22 | 50.57 | 48.71 | 18.38 | 18.06 |
| Middle point | (−90, −100) | (10, 15) | 21.01 | 24.09 | 18.02 | 17.93 |
| Edge point | <−105 | <5 | 12.66 | 10.18 | 7.92 | 6.24 |

*RSRP* Reference Signal Receiving Power
**SINR* Signal to Interference plus Noise Ratio

**Table 11.3** Coverage comparison (*unit* m)

| DL coverage | | UL coverage | |
|---|---|---|---|
| w/ | w/o | w/ | w/o |
| 600 | 607 | 598 | 607 |

Extensive test cases were performed to compare the system performance with and without the two technologies. In Tables 11.2 and 11.3, the coverage and throughput are compared. It can be seen that the performance difference is almost negligible after the adoption of both compression and SFBD.

Test results verified that compression (with a 2:1 compression ratio) and SFB are mature enough and the system performance is almost the same as without the adoption of the two technologies.

Despite the fact that combination of compression and SFBD can reduce fiber usage by 75 %, it is still far from enough for the large-scale deployment of C-RAN. From a future system upgrade perspective, it is obvious that capacity expansion and upgrade should not depend on the upgrade of fiber transport infrastructure since it would be too costly. Instead, it should rely on equipment updates. In this sense, WDM is a promising solution.

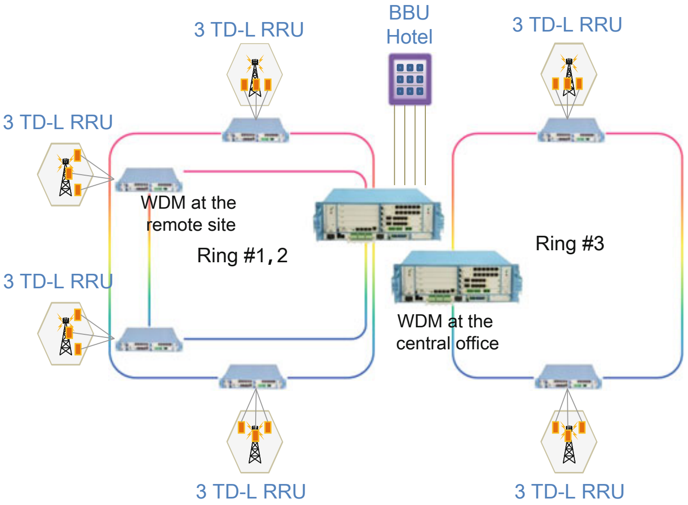


**Fig. 11.3** C-RAN field trial with WDM solution

The basic introduction of a WDM solution has been described in Sect. 11.4.2. So far in the industry, there have already been several vendors who claim to provide mature WDM fronthaul solutions. To further qualify the solution, based on the previous field trial in Fuzhou City (one of the three field trials for CPRI compression and SFBD), we reconstructed the C-RAN fronthaul utilizing WDM equipment. The network topology is shown in Fig. 11.3.

In this field trial, six remote WDM setups are installed at the remote RRU sites, each supporting one site, i.e., three LTE carriers. The BBU pool is connected with the local WDM equipment in the central office. Several WDM nodes are connected in a ring, with SFBD implemented. Due to power budget constraints, one such ring consists of only two WDM nodes. As a result, three fiber cores are needed for three rings of six WDM nodes. The number of supported LTE carriers is still 18, the same as in the previous trial. Compared with the previous test with compression and SFB which consumes 18 fiber cores (one fiber core per carrier), the reduction in the fiber consumption with WDM in this trial is fivefold.

The processing delay of the WDM nodes is empirically observed to be less than 1 μs, which is small enough to have no impact on CPRI transmission.

Another key feature of this WDM solution is protection switch (PS) capability. The operators' requirement in this case is less than 50 ms. In this field trial, both active and passive PSs are carefully examined. In active PS, we launched the switch command through a management system and found that the latency is only around 12 ms. Further, in the case of a passive PS, in which the fiber is pulled out to simulate link failure, the switch time is around 36 ms, meeting the requirement with room to spare.

To verify the maturity of the solution, we executed many O&M functions such as remote software download, remote system update, system backup and restoration, verifying that they all worked well.

In addition to the performance on the WDM network itself, we further tested the entire wireless system performance in terms of throughput, coverage, end-to-end latency, handover success rate and so on (the same tests as in the previous field trials for compression and SFBD). The key finding is that all the performance metrics are almost the same as in the previous trials. There is no impact with the introduction of WDM. Furthermore, the network continued to run for 3 months as a commercial trial network without failure.

## *11.5.2 Exploitation of C-RAN Virtualization*

### 11.5.2.1 Accelerator Design

In our first endeavor in C-RAN virtualization, we first developed a Proof of Concept (PoC) to evaluate the feasibility of baseband processing based on a GPP. There is further elaboration on this in [8]. Although the feasibility is verified, it is also found that almost 10 CPU cores are needed to process one 8-antenna 20-MHz TD-LTE carrier physical layer. Accordingly, the performance-to-power ratio of pure software BBU is relatively low. Based on these observations, we then conclude that a hardware accelerator should be used to speed up physical-layer processing so that more carriers can be handled with the same processing resources, and the performance-to-power ratio will be improved significantly. In this section, we introduce a design scheme for the accelerator. The design principles are as follows:

According to our evaluation and assessment on a preliminary pure software prototype, in DL, processing modules of high computation load include Fast Fourier Transformation (FFT), 8-antenna precoding and turbo encoding. The other modules, including scramble, modulation, and layer mapping, do not require high computation capability. They can be performed by GPPs as the latency is not critical. However, it is more reasonable to also incorporate these functions on the hardware accelerators from an interface data traffic point of view.

For each user's DL, Physical Downlink Shared Channel (PDSCH) processing not only requires a pending data block from the MAC layer of the user, but also needs the corresponding physical-layer configuration parameters of the user. The physical-layer DL processing modules on the hardware platform will process the MAC data on PDSCH according to the configured physical-layer parameters. On the other hand, the algorithms for bit-level processing and symbol-level signal processing of Physical Downlink Control Channel (PDCCH) are of low complexity and can be completed by a GPP. In order to call the physical-layer DL processing function modules, subcarrier mapping location information is required.

Similarly, in UL, processing modules of heavy computation are: Inverse Fast Fourier Transformation (IFFT), channel estimation, equalization, and turbo decoding of PDSCH. However, channel estimation is one of the most flexible parts in UL processing. Therefore, it is difficult to achieve direct compatibility among different vendors if it is realized in accelerators. The processing modules between equalization and turbo decoding are of low complexity, which can be performed by a general purpose CPU from the point of view of computing power and processing delay. However, it is more reasonable to process these parts on hardware accelerators from an interface data traffic point of view.

#### 11.5.2.2 Prototype Verification of Soft C-RAN

As mentioned in previous sections, virtualization is critical to realize C-RAN resource cloudification. Since virtualization is more an implementation issue than a research topic, the methodology we adopt to evaluate virtualization is to develop a PoC.

The prototype is developed based on standard IT servers of ×86 architecture. In this prototype, a dedicated hardware accelerator is used to process partial physical-layer functions including FFT, IFFT, channel coding, and decoding.

As shown in Fig. 11.4, all the physical servers are interconnected via three 10-Gbps Ethernet switches. They are used for carrying different traffic flows, including RT traffic, non-RT traffic, and virtualization OAM traffic. Another 1-Gbps switch is equipped for hardware OAM traffic. A standalone OAM server is provided for cloud management including virtualization management and hardware management.

There are two kinds of physical servers, i.e., modem servers and control servers, for hosting different applications. In the modem servers, TD-LTE L1 and TD-LTE L2 protocols are running in different VMs residing in different physical modem

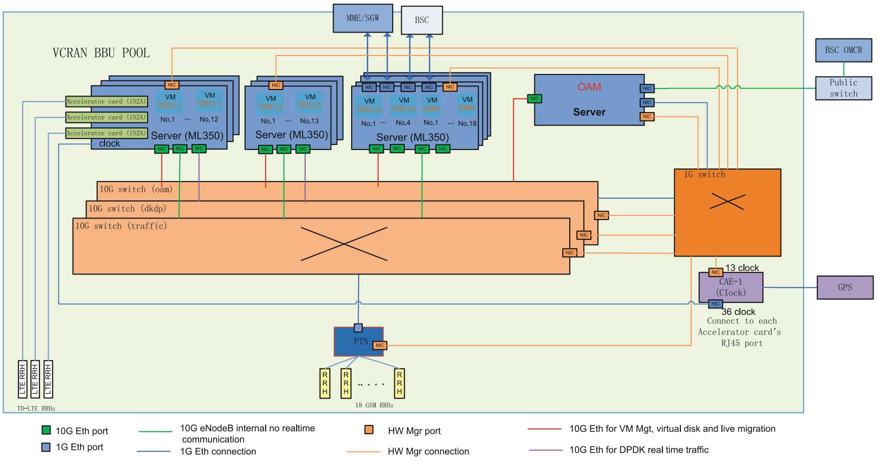


**Fig. 11.4** Architecture of virtualized C-RAN PoC

servers. In each modem server, several CPRI Peripheral Component Interconnect express (PCIe) plug-in cards are equipped for TD-LTE CPRI link termination. In the control servers, there are two kinds of VMs running, i.e., GSM VM and TD-LTE L3 VM. Control servers have external Ethernet connections to the EPC and the base station controller (BSC). Those external connections are used by the VMs for backhaul connection. For GSM VM, there are no dedicated accelerator cards. It is worth pointing out that in our demo except for the accelerator, all the other realizations in both GSM and LTE are based on commercial protocol stacks.

The prototype is first tested from a functional perspective. A GSM call is made between two GSM UEs, while an LTE terminal is simultaneously accessing the LTE VM and downloading files. The services ran well for at least 1 h.

In addition, we further evaluated the processing benchmark as well as interruption time performance with the following results:

- **Processing benchmark**

The processing benchmark for key L1 functions are calculated and shown in Table 11.4.

**Table 11.4** Benchmark for physical functions processing *(unit* μs)

| | | 20 MHz | |
|---|---|---|---|
| | | 2 Ant. (μs) | 8Ant. (μs) |
| *Uplink processing* | | | |
| PUSCH (single UE with UL 100 PRB)7 | 7.5 K shift + FFT | 105.9 | 189.9 |
| | Channel evaluation (IRC) | 147.7 | 524 |
| | PUSCH frequency domain processing | 52 | 218 |
| | MIMO/EQU | 179.8 | 384 |
| | IDFT | 34.6 | 34.6 |
| | Demodulation | 71.9 | 71.9 |
| | HARQ merge | 59.7 | 59.7 |
| | Turbo decode | 113 | 113 |
| | CRC | 9.6 | 9.6 |
| PUSCH | PUCCH format 1 per UE | 2.3 | 4.1 |
| | PUCCH format 2 per UE | 9.7 | 17.2 |
| *Downlink processing* | | | |
| PDSCH (single UE, 200 PRB, TM3/TM8, peak throughput) | CRC | 13.8 | 13.8 |
| | Turbo encode | 70 | 70 |
| | Scrambling | 14.9 | 14.9 |
| | Mod. | 36.4 | 36.4 |
| | Power control, precode, BF | 21 | 190 |
| | iFFT | 32 | 116 |

From the test results in Table 11.4, it can be calculated that on average 4 Sandy Bridge CPU cores are required for one 8-antenna cell with a specified level of code optimization for the commercial capacity requirement.

- **Evaluation of interruption time**

To further investigate the virtualization performance, the interruption time, which is a key indicator of the response time that a GPP system can have, is tested with the test results shown in Table 11.5. Both the internal interruption time triggered by the application and the external interruption time which is triggered by an external PCI card are tested. Thanks to the joint optimization on the virtualization systems such as the hypervisor and the OS, the interruption time is greatly reduced. It can be seen from the table that the maximum internal and external interruption times are 23 and 40 µs, respectively, while the averages are around 17 and 10 µs. By comparison, for a general Linux-based GPP platform, the typical interruption time ranges from 50 to 1000 µs. These results indicate that the system has the determinism required by hard RT applications. Considering that LTE processing usually has a 1 ms processing time budget, we then conclude that such latency is tolerable for LTE baseband processing.

## 11.6 Evolving Toward 5G

### *11.6.1 C-RAN to Enable Key 5G Technologies*

As described at the beginning of this chapter, in the 5G vision the network should evolve from cell-centric to user-centric, i.e., “No More Cell.” This idea means that users should be provided with not only much higher data rate services but also experience less difference between the cell-center and cell-edge regions. To achieve this goal, the severe interference experienced by the cell-edge users should be alleviated. Thus, coordination among multiple cells is necessary. In 5G networks, various coordination techniques have been proposed to solve the interference problem. However, algorithms such as Joint Transmission [2] cannot achieve maximum performance gain under the traditional architecture, e.g., LTE with an X2 interface which is of high latency and low bandwidth [13, 14]. C-RAN, on the other hand, thanks to the strong inherent central processing capability, provides an ideal structure to facilitate the implementation of coordination technologies with full or partial channel state information. In fact, it has been demonstrated that C-RAN can

**Table 11.5** Interruption time of the prototype (*unit* µs)

| | # of interrupts | Min | Max | Ave. |
|---|---|---|---|---|
| Internal interruption time | 95,309,564 | 15 | 23 | 17 |
| External interruption time | 217,587,258 | 5.59 | 40.01 | 10.19 |

greatly improve CoMP performance in terms of cell-edge spectrum efficiency by 119 % over non-cooperative transmission mechanisms [8].

C-RAN is also an ideal match for the deployment of an UDN which is deemed to be a promising solution for highly dense user traffic. The design of UDN involves joint consideration of many issues, including the control and user plane decoupling, inter-site carrier aggregation and coordination, and interference mitigation in a heterogeneous network. In this case, C-RAN can play an important role, with its internal high-speed low-latency switching mechanism and the central processing, for the implementation of those key technologies.

Finally, C-RAN provides a unique opportunity to support multi-RAT with the adoption of GPP and virtualization technology. In C-RAN, different RATs can be virtualized in the form of VMs and operate separately and independently on the same platform. Thanks to highly efficient VM communication, C-RAN can further help with multi-RAN coordination.

## *11.6.2 Rethink CPRI: CPRI Redefinition*

CPRI, the most common transmission protocol between BBU and RRU, has several disadvantages, which may make it not suitable for 5G networks:

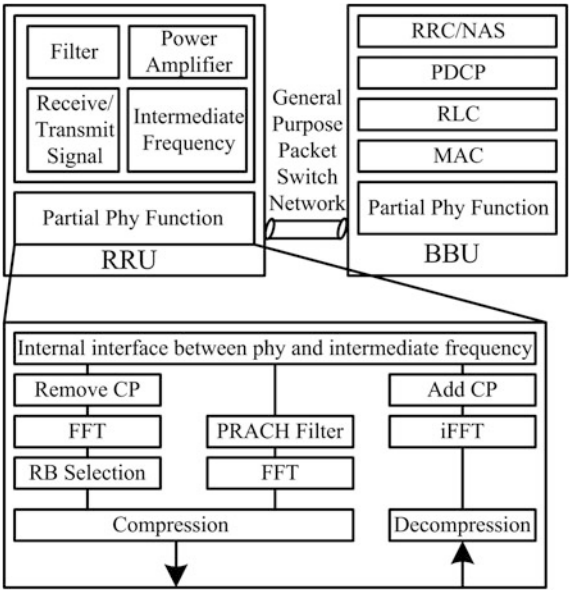
- The CPRI data rate is exceedingly high. One typical example is as described in Sect. 11.4.1; up to 9.8 Gbps is required for just one 20-MHz TD-LTE carrier with 8 antennas. It could be foreseen that in 5G, when large-scale antenna technology is introduced, the data rate will increase significantly.
- The CPRI data rate is constant, regardless of the real amount of user traffic. This makes the transmission extremely inefficient.
- The current CPRI link is point to point. However, in C-RAN with soft BBU pool, as pointed out before, there is no longer one-to-one correspondence between the BBU and RRU, and therefore a flexible fronthaul with multi-point to multi-point mapping capability is required.

CPRI redefinition aims at overcoming these shortcomings so as to have a better fronthaul interface for future 5G evolution. The basic idea of CPRI redefinition is to move a partial set of physical-layer functions to the RRU side and to realize packet transmission. Take LTE as an example. There are several possibilities on the function partition, and one example is shown in Fig. 11.5. In this scheme, the function set is comprised of the major PHY functions such as FFT/IFFT and Cyclic Prefix (CP) addition/removal. A new function block called Resource Block Selection (RB Selection) is implemented which only selects the scheduled (i.e., occupied) RBs to be placed into the compression module.

This scheme has several advantages:

(1) The data rate between the BBU and the RRU could be significantly reduced. As an example, an 8-antenna 20-MHz TD-LTE carrier without this scheme

**Fig. 11.5** Example of CPRI redefinition in LTE



has an I/Q bandwidth requirement of 9.8 Gbps. With this new scheme, the maximum bandwidth can be reduced to around 2.2 Gbps.

(2) CoMP implementation can still be supported since no CoMP-needed information is processed and terminated on the RRU.

(3) Due to the RB selection, the data rate between the BBU and the RRU is now elastic and varies with real user traffic, which is the opposite of the traditional case where the I/Q stream is constant even when there is no real traffic. This feature not only helps to reduce the capacity requirement on switching networks within the BBU pool but also reduces the switching latency. In addition, the data can now be encapsulated in the form of packets rather than a constant stream and therefore can be transmitted by a suitable packet switching protocol, such as Ethernet (which enjoys the benefits of improved flexibility and improved switching efficiency). Furthermore, statistical multiplexing gain can be used to improve the transmission efficiency.

The major problem with this idea is that it may require a major overhaul of existing standards. In addition, it will increase the complexity of RRU equipment due to the implementation of functions in the RRU. This in turn may make future system upgrades more difficult. Moreover, some key features such as support of CoMP may be lost. In the future, however, the idea of interface redefinition still becomes critical for some 5G technologies such as LSAS where there are a number of antennas on one RRU. It would be of interest to explore redesigning the BBU-RRU interface to achieve packetized fronthaul transmission while still keeping the RRU as simple as possible.

### 11.6.3 Edge Application on C-RAN

With the large-scale commercialization of LTE networks, a wide variety of bandwidth-hungry and latency-sensitive mobile data applications are expected to emerge. For operators, the profit margins of traditional telecom services (e.g., voice and SMS) have been going down continuously. Despite the heavy investment of operators in network construction and maintenance, mobile networks tend to become “dumb pipes” of over-the-top (OTT) Internet companies. In addition, self-operated data services are also facing fierce, homogeneous competition, which makes the existing mobile business model hard to sustain in the near future. It is critical for operators not only to reduce network expansion cost, but also to provide differentiated QoE for mobile subscribers.

On the other hand, mobile base stations, as operators’ important and differentiated assets, have not been utilized to their full capacity. The idea of combining C-RAN with applications is to exploit their unique advantage by building applications over the edge of mobile networks, especially the C-RAN BBU pool. It is expected that, in this way, operators can take full advantage of distributed computation and storage capabilities to reduce network congestion and latency and also provide differentiated QoE to improve subscriber loyalty.

As shown in Fig. 11.6, edge applications and BBU pool software are deployed over the same hardware platform and isolated from each other via VMs. It is expected not only to reduce backbone traffic and latency, but also to provide rapid

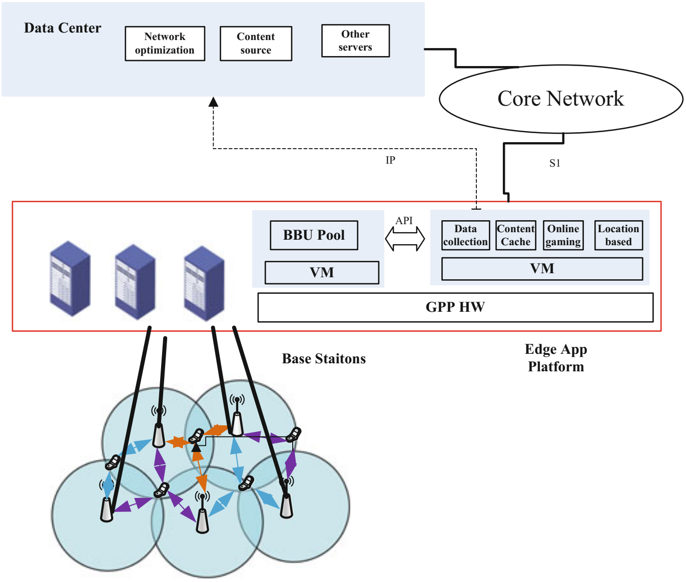


**Fig. 11.6** Edge applications architecture over C-RAN GPP BBU pool

time-to-market deployment via GPP platforms and virtualization technologies. The architecture has the following characteristics:

(1) **Distributed computation and storage**: User plane traffic can be locally cached and processed. If we deploy applications over a C-RAN BBU pool serving 5000–1000 users, the hit rate of content retrieval can be guaranteed and deployment costs can be reduced as well.
(2) **Radio API**: By providing real-time and refined radio network information, e.g., RT loading and radio link status, applications on edge can be improved further.
(3) **GPP platform and IT technologies**: Low cost and faster development, release, and maintenance can be achieved with mature development toolkits when using GPP platforms.
(4) **Collaboration with cloud data centers**: Applications in cloud data centers and over the edge can collaborate and complement each other to build a smarter pipe.

## 11.7 Conclusions

Traditional network design is cell-centric, which has become more and more unsuitable for future 5G systems, since it fails to take into account traffic variation and diverse environments. It is well recognized that 5G should be designed starting with a paradigm shift from cell-centric design toward "*No More Cells*," i.e., user-centric design in order to better deal with tidal effects and improve system efficiency. The realization of "No More Cells" can be facilitated by C-RAN which virtualizes the processing resources in a pool and is capable of dynamically allocating the resources on demand.

C-RAN is a revolutionary evolution of radio access network. It is not only a key enabling element of future 5G systems but also offers a significant enhancement path to all existing systems including 2G, 3G, and 4G. In this chapter, a comprehensive study of C-RAN was provided. Due to the inherent high-speed, low-latency nature within the C-RAN resource pool, different information such as the channel state information and scheduling information can be shared among different nodes in a timely manner. Consequently, key enabling technologies toward user-centric networks, such as CoMP, can be realized efficiently.

However, two major kinds of challenges need to be addressed for C-RAN realization: an efficient fronthaul solution as well as real-time virtualization implementation. For the former challenge, we demonstrated through various field trials different approaches, including CPRI compression, single-fiber bi-direction, and WDM. It is verified that WDM-based fronthaul significantly reduces fiber consumption, supports flexible topology, and facilitates future system upgrade along with mature management capability. In one of the field trials, only three fiber cores are used to aggregate 6 sites of 18 TD-LTE carriers. Thus, existing WDM solutions are mature enough to enable large-scale C-RAN deployment in existing systems such as 4G.

On the road toward C-RAN cloudification, it is found that an accelerator is still currently necessary. Moreover, virtualization implementation imposes several challenges, including optimization on hypervisor and OSs, management functions, and I/O virtualization. A prototype with commercial LTE protocol stack implemented on GPP was developed. It successfully demonstrated multi-RAT support, i.e., GSM voice calls and LTE data services, as well as satisfactory performance in interruption latency. Further evaluation showed that 4 Sandy Bridge cores are needed to process a 20-MHz TD-LTE carrier with 8 antennas.

In the long run, C-RAN will not only facilitate the realization of “No More Cells” for a user-centric 5G system, but also remain an integral part of future greener and softer systems.

**Acknowledgments** We would like to express our sincere gratitude to all of our partners, including Alcatel-Lucent, Intel, and Wind River for the development of the GPP-based C-RAN prototype. We would also like to thank all C-RAN and 5G team members in China Mobile for their helpful discussion and valuable comments.

## References

1. Chih-Lin I, Rowell C, Han S, Xu Z, Li G, Pan Z (2014) Toward green and soft: a 5G perspective. IEEE Commun Mag 52(2):66–73
2. 3GPP TR 36.819 (2011) Coordinated multi-point operation for LTE physical layer aspects (Release 11). Version 11.1.0
3. http://www.greentouch.org
4. Kishiyama Y, Benjebbour A, Nakamura T, Ishii H (2013) Future steps of LTE-A: evolution toward integration of local area and wide area systems. IEEE Wirel Commun 20(1):12–18
5. Chen Y, Li G, Cui C (2014) Macro assisted ultra-lean data carrier and architectural design. ZTE Technol J 2:17–21
6. Marzetta T (2010) Noncooperative cellular wireless with unlimited numbers of base station antennas. IEEE Trans Wirel Commun 9(11):3590–3600
7. Wu J, Rangan S, Zhang H (2013) Green communication. CRC Press
8. C. M. R. Institute (2013) C-RAN white paper 3.0: the road towards green ran. http://labs.chinamobile.com/cran
9. http://www.cpri.info/
10. Sezer S et al (2013) Are we ready for SDN? Implementation challenges for software-defined networks. IEEE Commun Mag 51(7):36–43
11. ETSI NFV ISG (2012) Network functions virtualisation. http://portal.etsi.org/portal/server.pt/community/NFV/367

12. Younge AJ et al (2011) Analysis of virtualization technologies for high performance computing environments. In: IEEE international conference on cloud computing. Washington DC, USA, pp 9–16
13. CMCC. Simulation results for CoMP Phase I (2011) Evaluation in homogeneous network. R1-111301, 3GPP TSG-RAN WG1 #65, Barcelona, Spain
14. Wang QX, Jiang DJ, Liu GY, Yan ZG (2009) Coordinated multiple points transmission for LTE-advanced systems. In: Proceedings of 5th international conference on wireless communications, networking and mobile computing, pp 1–5