SIP (2017), vol. 6, e10, page 1 of 16 © The Authors, 2017.

doi:10.1017/ATSIP.2017.11

INDUSTRIAL TECHNOLOGY ADVANCES

Seven fundamental rethinking for next-generation wireless communications

CHIH-LIN I

The fifth-generation (5G) mobile communication networks, which are anticipated to be soft, green, and super-fast, may possibly be deployed in 2020s to satisfy the challenging demands of mobile communication in various scenarios. Characterized by a mixed set of key performance indicators like data rates, latency, mobility, energy efficiency, and traffic density, 5G services demand a fundamental revolution on the end to end network architecture and key technologies design. Toward a "soft, green, and super-fast" 5G, this paper presents seven innovative 5G R&D themes of China Mobile, including: (1) rethinking Shannon to start a green journey on wireless systems; (2) rethinking Ring and Young for no more "cells"; (3) rethinking signaling and control to make network applications aware and load aware; (4) rethinking antennas to make base stations invisible via SmarTiles; (5) rethinking spectrum and air interface to enable wireless signals to "dress for the occasion"; (6) rethinking fronthaul (FH) to enable Soft RAN via next-generation FH interface; and (7) rethinking the protocol stack for flexible configurations of diversified access points and optimal baseband function split between the base band unit pool and the Remote Radio Systems.

Keywords: 5G, User-centric network, Software-defined air interface.

Received 5 May 2016; Revised 9 August 2017

I. INTRODUCTION

With the global commercialization of the fourth-generation (4G) long-term evolution (LTE) standard, the wireless community is now looking forward to the next-generation mobile network. According to International Telecommunication Union, the official name of the next-generation mobile network is International Mobile Telecommunications (IMT)-2020, hereafter termed 5G for short, which will be launched in around 2020. Worldwide 5G R&D has been extensively carried out, starting with an investigation on user demands, on application scenarios, and on technical trends [1–4]. Quite recently, the campaign on 5G standards has just begun, with diversified proposals on timeline, work scope, key technologies, and spectrum strategy being intensively discussed.

In 5G era, mobile Internet and Internet of things are the two main drivers for 5G services. The 5G scenarios include at least dense residential areas, office towers, stadiums, open-air gatherings, subways, highways, high-speed railways, and wide-area coverage. These scenarios, which are characterized by ultra-high traffic volume density, ultra-high connection density, or ultra-high mobility, raise extreme challenges for 5G. Typical services, such as augmented reality, virtual reality, ultra-high-definition video, cloud storage, Internet of vehicles, smart home, and over-the-top services, will be provided in these scenarios.

The performance requirements for 5G are derived for each scenario, according to the predicted distribution of users, percentage of different services, and service requirements such as data rate and latency. The key performance indicators (KPIs) for 5G include user experienced data rate, connection density, end-to-end latency, traffic volume density, mobility, and peak data rate. The KPIs proposed by China's IMT-2020 Promotion Group include, e.g. over 100 Mbps user experienced data rate, one million connections per square kilometer, 1 ms end-to-end latency and tens of Gbps peak data rate [4]. To meet the extremely challenging user demands driven by mobile Internet and Internet of things in a highly efficient way, 5G networks are anticipated to be *soft*, *green*, and *super-fast*.

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

Green Communication Research Center, China Mobile Research Institute, Beijing 100053, People's Republic of China

Corresponding author:
Chih-Lin I
Email: icl@chinamobile.com

operation, and management. In a soft network, the computing, storage, and radio resources are virtualized and centralized to achieve dynamic and user-centric resource management, matching service features. Soft networks are expected to build on a telecom-level cloud platform to enable network-as-a-service with the features of open network capability and network sharing. This makes it possible to achieve network flexibility and scalability and provides users with massive variety of services and consistent quality of experience (QoE). Soft networks may achieve breakthroughs first in cloud-radio access network (C-RAN) [5], network function virtualization [6], and software defined network (SDN) [7] with control and data decoupling.

The soft network concept should be extended to the air interface as well. Instead of a global optimized air interface which is a trade-off among many factors, a software-defined air interface (SDAI) will be considered, where the air interface can be optimized to each individual application scenario via flexible configuration of spectrum, bandwidth, waveform, duplex, and multiple access schemes, etc. This enables broad adaption of future networks to application scenarios with extreme diverse requirements.

Green communication is a social responsibility to reduce energy consumption as well as an economic target for wireless communication industry. High spectrum, spatial, temporal, hardware, software resource efficiency, low-power consumption, and low cost are the basic requirements of a green 5G network. Green networks will achieve a 1000-fold capacity increase with minimum burden of spectrum resources. Advanced signal processing to effectively explore spatial resources, centralized coordination to reverse harmful interference to the useful signal, joint baseband and radio frequency (RF) processing to enhance the same spectrum duplexing, etc. are some of the key technologies to improve radio resource efficiencies.

Green networks will achieve 100 times energy efficiency (EE) improvement to reduce operating expense for sustainable operations. It requires a capability for end-to-end energy management and optimization, so that the total energy consumption will be minimized while meeting service requirements. Green networks enable network capacity migration and breathing to match service variations without a waste of network resources. Moreover, "plug and play" and on-off nodes are also essential parts of a green network. These massive nodes work without network planning in advance. Thus, an advanced self-organizing network is actually important for dynamic network planning and topology, as well as near real-time network optimization. Green networks are able to utilize renewable energy, such as wind and/or solar energy as alternative power supply for networks, and bioelectric, kinetic, and/or thermal energy for terminals.

Super-fast 5G network is anticipated to provide fiber-like access data rate, "zero" latency user experience, and ultra-high mobility, and is envisioned to approach immersive and tactile user experience in any extreme scenarios. An immersive user experience can be achieved with further development of mobile Internet with high-definition video-dominated applications. To this end, a 1000-times greater network capacity is expected by 2020 with 20 and 10 Gbps peak data rate requirements for DL and UL, respectively. Further exploration in spatial domain, wideband systems with up to 500 MHz bandwidth in higher frequency, multi-connection in ultra-dense network (UDN) scenario, and other areas will be considered.

Use case scenarios such as remote surgery, auto-pilot, and on-line gaming need a tactile round trip response. An end-to-end latency smaller than 10 ms is expected for future network with a smaller than 1 ms delay budget reserved for air interface. New frame structures and access scheme based on new waveform design should be pursued with this target. 5G should cover mobility up to 500 km/h, due to the wide deployment of high-speed trains in China.

Toward *soft*, *green*, and *super-fast* 5G networks, in this paper, the design methodologies are presented, from the perspective of China Mobile. The main 5G R&D themes are elaborated via seven fundamental rethinking, including rethink Shannon in Section II, rethinking Ring and Young in Section III, rethink signaling and control in Section IV, rethink antennas in Section V, rethink spectrum and air interface in Section VI, rethink protocol stack Section VII, and rethink fronthual in Section VIII. The paper is summarized in Section IX.

II. RETHINK SHANNON

After decades of high-speed development, the scale of information and communication technology, or particularly communication networks, is huge enough such that its power consumption is no longer a negligible factor in global energy consumption. Considering 1000 times capacity increase by the year 2020, the power consumption of future networks is not affordable if the network is designed with the current energy scaling rule.

Given limited spectrum and ever-increasing capacity demand, spectrum efficiency (SE) has been pursued for decades as the top design priority of all major wireless standards, ranging from cellular networks to local and personal area networks. The cellular data rate has been improved from kilobits per second in 2G to gigabits per second in 4G. SE-oriented designs, however, have overlooked the issues of infrastructure power consumption. Currently, RANs consume 70% of the total power. In contrast to the exponential growth of traffic volume on mobile Internet, both the associated revenue growth and the network EE improvement lag by orders of magnitude. A sustainable future wireless network must therefore be not only spectrum-efficient but also energy-efficient. Therefore, EE and SE joint optimization is a critical part of 5G research [8–10]. Looking at traditional cellular systems, there are many opportunities to become greener, from equipment level such as more efficient power amplifiers (PAs) using envelop tracking, to network level such as dynamic operation in line with traffic variations both in time and space. For fundamental principles of EE and SE co-design, one must first revisit the classic Shannon

theory and reformulate it in terms of EE and SE. In classic Shannon theory, the channel capacity is a function of the log of the transmit power (P_t), the noise power spectral density (N_0), and the system bandwidth (W). The total system power consumption is a sum of P_t and the circuit power P_c (the power consumption in the base station (BS) which does not scale with the transmit power P_t),

$$P_{tot} = P_t/\rho + P_c, \quad (1)$$

where ρ is PA efficiency defined as the ratio of the input of the PA to the output of the PA. From the definition of EE, EE is equal to the channel capacity normalized by the system power consumption. SE is the channel capacity normalized by system bandwidth. The relationship of EE (η_{EE}) and SE (η_{SE}) can be shown as a function of PA efficiency and P_c in Fig. 1. It can be observed that when P_c is zero, there is a monotonic trade-off between η_{EE} and η_{SE} as predicted by the classic Shannon theory. For non-zero P_c, however, η_{EE} increases in the low SE region and decreases in the high SE region with η_{SE} (for a given η_{EE}, there are two values of η_{SE}). As P_c increases, the EE–SE curve appears flatter. Furthermore, when taking the derivative of η_{EE} over η_{SE}, the maximum EE (η^*_{EE}) and its corresponding SE (η^*_{SE}) then satisfy the following:

$$\log_2 \eta^*_{EE} = \frac{\log_2 \rho}{N_0 \ln 2} - \eta^*_{SE}. \quad (2)$$

This means there is a linear relationship between $\log_2 \eta^*_{EE}$ and η^*_{SE}. Similar to the EE–SE relationship with classic Shannon theory, a higher η^*_{SE} will always lead to a lower η^*_{EE}. However, the EE–SE relationship at the EE optimal points is independent of P_c. This observation implies that as P_c decreases, an exponential EE gain may be obtained at the cost of linear SE loss.

As explained in [8–10], the parameters affecting EE and SE trade-off include actually all the parameters of the system, e.g. P_c, η_{EE}, η_{SE}, antenna number, system bandwidth,

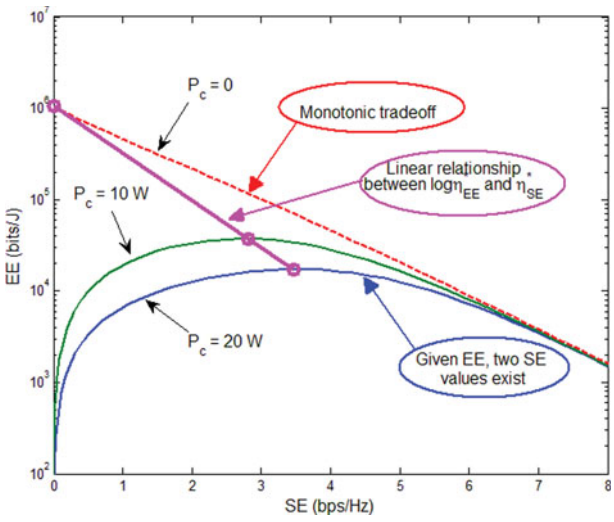

Fig. 1. SE and EE relationship for different circuit powers.

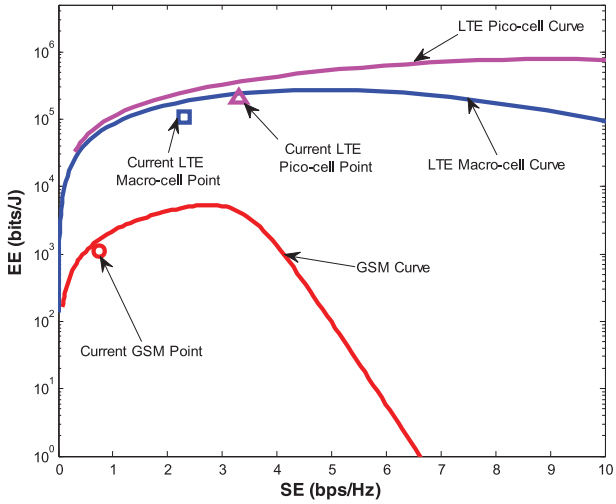

Fig. 2. SE and EE relationship for current cellular networks.

ρ, number of data streams, etc. The practical meaning of this analysis is that the wireless system need to operate at exactly the maximum EE point, with the corresponding SE high enough to meet the system requirement. To reach this goal, all the system parameters need to be designed jointly.

Figure 2 compares the EE–SE performance of current Global System for Mobile Communications (GSM) and LTE BSs [1]. LTE performs better than GSM in terms of both SE and EE; both, however, are working in a low SE region, indicating room for improvement. As can be seen, the current LTE and GSM EE–SE points are below the theoretical curves. The intuition behind the possibility of both systems to enhance the EE–SE performance is that the PA efficiency is not high, generally 30–40%, and the circuit power is rather high, which generally includes the power consumption of cooling at the BSs. A straightforward way to improve EE–SE is via central processing at the C-RAN architecture, which dramatically reduces the power consumption of cooling at millions of BSs.

The SE and EE trade-off can be applied to various design aspects in 5G. For example, when a large number of antennas are implemented to achieve better beamforming gains, implementing the same number of transceivers may not be feasible due to excessive demand on real-time signal processing for high BF gains, high power consumption and cost (especially the high cost and power consumption of mixed-signal devices in millimeter wave (mmW) systems). The beamforming structure with much smaller number of digital transceivers than total antenna number will therefore be more practical and cost-effective to deploy. The EE and SE trade-off was investigated in [10].

III. RETHINK RING AND YOUNG

The concept of cellular systems was proposed in 1947 by two researchers from Bell Labs, Douglas H. Ring and W. Rae Young. Since the first generation of cellular standards, this cell-centric design has been maintained through every

new generation of standards including 4G. Toward the timeline of 2020 with the introduction of Heterogeneous Network (HetNet) and UDN [11], multiple layers of radio network have come into being. Energy consumption, interference, mobility issues are becoming more serious due to smaller inter-site distance. Diverse types of BSs with different coverage, transmit power, and frequency bands tend to be introduced. Traffic fluctuation is more significant than before, taking into account emerging millions of mobile data applications. Even different types of radio interfaces may be introduced to handle highly diversified requirements [12] from enhance board band (eMBB), ultra-reliable low latency communication (URLLC), and massive machine-type communication (mMTC) in 5G era. Therefore, in practical deployments, it is clear that the traditional homogeneous cell-centric design of mobile network does not match the anticipated traffic variations and diverse radio environments.

Moreover, for the traditional definition of RAN, a user equipment (UE) is managed within a cell of a BS, which means that the context of the UE (e.g. radio bearer configuration, UE ID, radio resource allocated) is directly bundled to a cell instead of a BS, and then the cell is allocated within the BS. The cell-based design faces many challenges, such as less flexible control, higher latency signaling procedure, and more complex processing inside BS, which leads to difficulty in fulfilling the requirements of 5G.

The design of user-centric 5G radio networks should start with the principle of "No More Cells", departing from cell-based coverage, resource management, and signal processing.

A) New definition of "UE" and "cell" in 5G RAN

In order to achieve low latency and robust data link, high-performance control procedure and real-time flexible control over air interface, decoupling UEs from "cell" is required. Eventually, "UE" and "cell" are in parallel both the fundamental elements of 5G networks, and "cell" becomes one of the dedicated radio resources. In order to achieve this kind of new relation between UE and cell, it seems proper to separate the UE management and cell management in the BS. For example, we may introduce the UE Management Function and the Cell Management Function.

UE Management Function is responsible for the management of the context, data, and all kinds of resources allocated to UEs. The content and the resources of UE can be obtained from the Cell Management Function. Cell Management Function is responsible for the management of all common resources, which are not allocated to any UE.

B) Benefits of new definition of "UE" and "cell" in 5G RAN

For dual connectivity defined in 3GPP Rel-12, UE and cell are relatively tightly coupled, and thus UE with related context is bundled with Master eNB (MeNB). As a result, this is not flexible and may bring excessive signaling overhead by taking in account the possibility of promoting Secondary eNB (SeNB) to MeNB due to UE mobility. And furthermore, MeNB takes most of Signaling Radio Bearer (SRB) processing and considerable amount of Data Radio Bearer (DRB) processing, e.g. Packet Data Convergence Protocol (PDCP) split bearer flow control, aggregation, reordering, etc., which may cause unbalanced load distribution between MeNB and SeNB.

Instead, the UE data and contexts are managed by UE Management Function instead of a specific cell (e.g. primary cell, PCell) or a specific eNB (e.g. MeNB). During the data transmission over air interface through multiple "severing" cells, all the data and UE contexts can be directly provisioned by the UE Management Function; therefore, there is no need for the scheduled cell to acquire the data and UE contexts from other severing cells. This function design leads to a wide room to establish a real-time and flexible radio resource allocation mechanism in 5G RAN, which brings a significant advantage in efficient data processing and optimized system performance in the context of multiple cell operation.

Each UE of the UE Management Function can be flexibly allocated with one or multiple cells belonging to the Cell Management Function, and on the contrast each cell of the Cell Management Function can also flexibly allocate resource to UEs. Based on the new relation between the UE Management Function and the Cell Management Function, the number of critical nodes for data links and control links reduce from three levels to two levels (BS → cell → UE to BS → UE).

C) New potential RAN architecture

With new definition of "UE" and "cell", UE-level context, control, and management is required to be centralized to achieve low latency and robust data link. Given a great deal of overlapped coverage in dense deployment, more centralized collaboration may be needed to alleviate interference, corporative scheduling and transmission, improve mobility robustness, etc.

Naturally, a potential architecture for next-generation RAN is illustrated in Fig. 3. The new BS may consist of Radio Cloud Center (RCC) and Remote Radio System (RRS) nodes as shown in Fig. 3. RCC is a central unit (CU), which may incorporate higher layer control and data functions, and possibly some baseband functions as well. RCC would be able to handle multiple cells and serves as a function pool. RRS are distributed remote unit, which could implement the RF part (i.e. remote radio head, RRH) and the remaining baseband processing, which depends on different options of function split by taking into account diverse profiles of fronthaul (FH) network and RAN deployment (e.g. number of antennas).

- If the FH is ideal enough without restriction of latency and transmission bandwidth, e.g. limited number of antennas (like 2,8) with dark fiber, base band unit (BBU) and

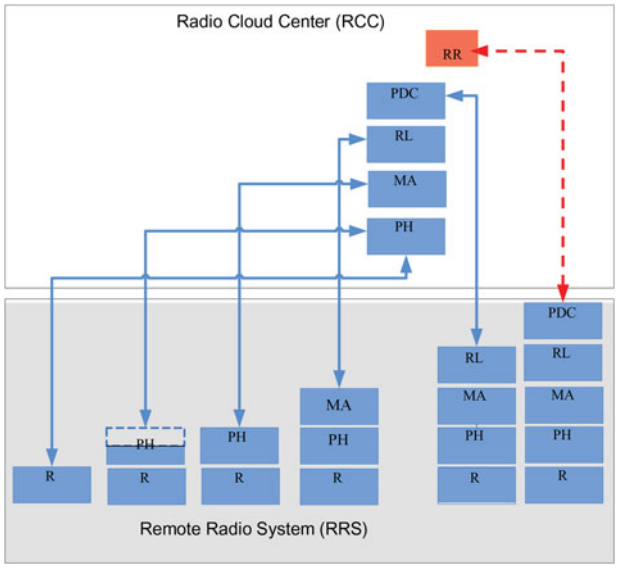

Fig. 3. New RAN architecture and examples of functional split options. Note: Control plane: red dotted; user plane: blue solid.

RRH separation can be considered to achieve maximum collaboration benefits.

- If a large number of antennas (like 128 and even more) are deployed, function split of internal physical or PHY and media access control (MAC) layer may be preferred, so that fronthauling of antenna-level data processing can be avoided.
- If the FH is not that ideal with restriction of latency and transmission bandwidth, PDCP or radio link control (RLC) level functional split can be considered.

On the other way round, though the "connection-oriented" signaling is perfectly suitable to the most resource consuming video services in 3G/4G or even 5G for connection setup, there is a challenge as how to optimize the rate of the connection under the resource contending radio environment, and to improve user satisfaction.

Rethinking signaling and control is to break through the conventional "one-fits-all" network architecture and procedures, and to make the network become context aware and service customized. It is to serve the different requirements with high efficiency, to optimize diversified user experiences under the resource contending radio environment. It is proposed that the 5G signaling/control must be application aware, load condition aware, and user status (e.g. mobility) aware. On the one hand, 5G over-the-air signaling must be an intelligent combination of both connection-oriented and connectionless mechanisms; on the other hand, the mobile networks shall be capable to provide on demand and customized network functions for differentiated user and traffic characteristics, e.g. mobility management, access, and scheduling.

High layer split (PDCP and RLC) has been agreed in 3GPP for standardization. The interface between protocol stacks above PDCP in CU and protocol stacks including RLC and below in DU (distributed unit) will be defined to allow for interoperability. Low layer split options have also been agreed for further study in 3GPP.

D) Adaptive multiple connections

In 5G era, multi-connection is an inherent key feature of 5G RAN including decoupled control and user plane, multiple user plane data links, decoupled DL and UL, which aim to enhance coverage, mobility, EE, and spectral efficiency.

With decoupled signaling and data, the mobility robustness can be easily improved since handover signaling overhead is reduced with a more stable signaling connection with macro signaling BS, while the small-cell deployment becomes much easier since no careful cell planning is required anymore. Spectrum utilization in small cells will be significantly enhanced due to the much relaxed requirement of control information and reference signals transmission from small cells. The control information can be transmitted from either the macro BSs or small cells. For fast moving users, both DL and UL control connection with the macro cell can efficiently eliminate the possibility of frequent handover, thus significantly reducing the related signaling overhead. On the contrary, it is suggested that the control link can be established with small cells for slowly moving users.

By decoupling of the DL and UL, flexible resource allocation can be facilitated between cells. In the traditional cell-centric network, the DL and UL connections are established with the same BS. However, in HetNet deployment, the nearby small cells with fewer DL reference signal power may possibly provide better UL connection. Therefore, the DL and UL of one UE may well be established with different BSs. Global resource optimization in user-centric design involves optimal selection of DL and UL connections for both control and data flows of all users. This optimal multi-connection issue is not feasible in traditional RAN, because too much inter-BS information sharing will be incurred, including dynamic user channel state information and scheduling information, etc.

In the user-centric HetNet with decoupled control and data, decoupled DL and UL, any information (control or data) can be flexibly transmitted to each user from one or multiple points. The optimal transmission point selection needs to consider the traffic load of each point, quality of service (QoS) or QoE, user's mobility status, energy consumption of transferring of the related information, channel state information, and induced signaling overhead.

Contrary to the traditional communication systems, the network topology of UDN is quite complex. In some scenarios, a single control connection with the nearest small cell may not be adequate, e.g. the DL signal quality is not good enough. Multi-connection is hence motivated. More than two nearby small cells can be accessed, and maintain control channels to the same user. Note that mobility support is better provisioned via multi-connection mechanism, which generally will incur redundant signaling overhead in establishing and maintaining more than one connection simultaneously. Therefore, the multi-connection of control

channels should follow some trigger mechanism, e.g. if the signal quality of the strongest DL control channel is not satisfactory, an alternative control link can be established with another adjacent small cell.

The capability of control and data decoupling in HetNet has phenomenal impacts on the design of small cells. For example, the cell common control information can only be transmitted from the macro cell. This indicates that small cells can be simplified and designed in extreme case, e.g. cell-specific reference signal and system information may not be needed for small cells.

Multi-connectivity is being discussed in 3GPP as a key feature of 5G. Inter-RAT multi-connectivity between LTE and new radio (NR) is required to take advantage of LTE continuous coverage with boosted capacity of NR access. In addition, more enhancements will be introduced, e.g. data duplication to guarantee ultra-reliability and enhanced flow control to implement higher SE.

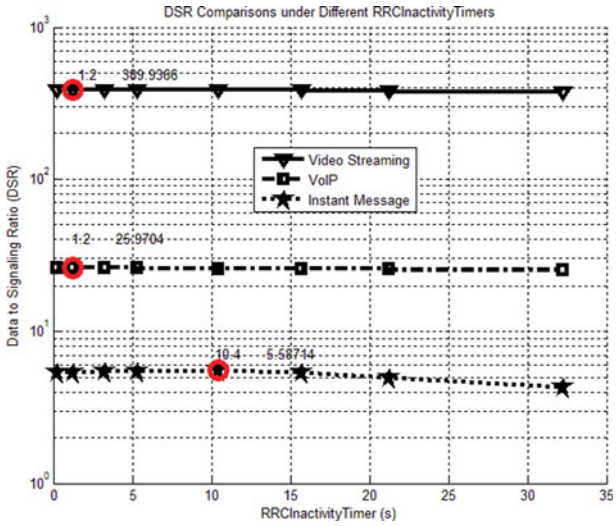

Fig. 4. DSR comparison of IM to video and VoIP under different RRC Inactivity Timers.

IV. RETHINK SIGNALING AND CONTROL

As it is known, the existing cellular network is connection oriented. A standard connection should always be built before any data transmission. Two radio resource control (RRC) states are defined for the management of different users, RRC_IDLE for data inactivity and RRC_CONNECTED for data transmission. By setting an RRC connection, a user enters RRC_CONNECTED state from RRC_IDLE state. On the other hand, the user enters RRC_IDLE state when the connection is released. Normally, release of a connection is triggered if the user is inactive for a certain timing period of an RRC Inactivity Timer. Then the user needs to re-establish an RRC connection to continue to transmit data. A typical RRC connection establishment/release process involves more than 12 interactions in RAN side and 15 interactions in CN side. It is prerequisite to synchronize user context on different network entities. Once the user is in RRC_CONNECTED, connection maintenance is required, including channel quality feedback, sounding signal transmission, and handover between cells, etc. Apparently, such "connection-oriented" procedures cannot satisfy service of critical performance parameters, e.g. ultra-low latency, as the procedures take at least tens of millisecond. Moreover, such "connection-oriented" signaling can be extremely inefficient if applied to some emerging services, e.g. the burst-type data like instant message (IM). As illustrated in Fig. 4, small-data bursts had exhibited orders of magnitude higher over-the-air signaling overhead than more traditional streaming services, by metric of ratio of pure data bits to corresponding supporting signaling bits (DSR) [13].

This is because most small-data-dominated applications, like IM, generate a constant stream of autonomous traffic all the time, erasing the previously clear demarcation between data activity and data inactivity. Thus, user transfers between RRC_CONNECTED and RRC_IDLE state frequently, resulting in large RRC signaling. Though effort has been made to reach optimal balance of RRC transitions and cost of RRC_CONNECTED state maintenance with traffic- and mobility-aware RRC Inactivity Timer, DSR efficiency is still low for the sporadic small-volume data, especially for keep-alive (KA) messages of mobile Internet traffic. Since for the data, it is both costly to maintain an RRC connection and to re-establish an RRC connection.

In 3GPP 5G standard, a concept named "network slicing" is introduced as a 5G key feature. This new feature is to support the diverse 5G applications and scenarios with flexibility and scalability. It is consistent with our thought of rethinking signaling control and to enable high-efficiency, service-customized network architecture and protocols. Meanwhile, conventional end-to-end QoS framework is replaced by finer grain QoS flow and RAN-based QoS flow-to-DRB mapping. In this way, tailored QoS can be decided by RAN rather than CN policy. RAN can do service scenario-customized QoS control. This new QoS concept is also a good example of our rethinking signaling and control concept. In following subsections, more use cases of "rethinking signaling and control" in RAN are illustrated.

A) RRC optimization for small data

As described, a UE synchronizes with the cellular network both in the DL and UL, establishes an RRC connection, and then enters RRC_CONNECTED state to acquire data transmission capability; but as have been mentioned, cost of RRC connection establishment can be extremely high for small data. Yet, it is the premise for any data transmission. So it is proposed in this section that a slim RRC state be introduced to support low signaling/control overhead for small-data transmission; it is named RRC_KEEP_ALIVE state. This RRC_KEEP_ALIVE is characterized as follows [13]:

First, RRC_KEEP_ALIVE supports transfer of small data to/from UE with customized slim signaling. For example, by employing a small-data indication in RRC connection

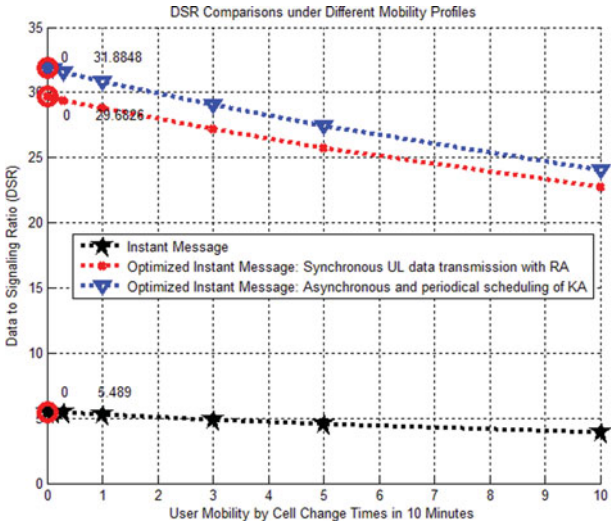

Fig. 5. Optimized DSR performance under different mobility conditions[1].

request, customized slim signaling can be provisioned for small-data transmission. Design of slim signaling can refer to control plane-based transmission and user plane transmission approaches of narrow band – Internet of things.

Second, RRC_KEEP_ALIVE does not require handover behaviors when user moves across cells. Instead, when cell transfer happens, the source cell simply releases user context on expiry of a timer without explicit signaling exchanges; or only when data transfer happens in another cell, context maintained in source cell can be transferred to target cell. This can be done because small data are usually sporadic and not continuous, and they do not bother to transfer the context. On the other hand, to enable faster connection, e.g. for URLLC services, the user context can be preserved or transferred without explicit signaling and activated fast by new data request.

Third, RRC_KEEP_ALIVE does not require performing periodic channel quality indicator, sounding reference signaling, and inter-/intra-frequency measurement, since continuous channel estimation and feedback is not necessary for sporadic small-data transmission.

As illustrated in Fig. 5, DSR performance of IM is improved by sixfold with the introduction of optimized RRC state together with the optimized signaling flows. The proposed RRC_KEEP_ALIVE state mitigates impact of RRC maintenance signaling overhead, and effectively supports slim RRC setup signaling, which therefore enables great DSR gain.

In 3GPP NR, a new RRC protocol state RRC_INACTIVE is introduced. This RRC_INACTIVE mode is very similar to our proposed RRC_KEEP_ALIVE state in following aspects: First, in RRC_INACTIVE, a connection is established for UE between the core and RAN, and UE access stratum context is stored in at least one gNodeB and the UE; second, UE in RRC_INACTIVE behaves like an ilde UE. It does cell reselection mobility rather than handover mobility; third, it would be further discussed in late Rel15 or Rel16, the possibility of small-data transmission in RRC_INACTIVE.

[1]Synchronous UL data transmission with RA, is uplink data transmission after uplink radio interface synchronization by random access; asynchronous and periodical scheduling of KA, is uplink KA message transmission with periodical resource allocation according to KA periodicity and without random access procedures [13].

B) Cross-layer optimization for video service

Operators are facing more and more challenges in providing mobile broadband services due to high-data rate and low-latency requirements and also due to the following constraints: first, isolated design of the radio network, e.g. radio resource scheduling based on varying radio channel and the applications, e.g. adjustments of video coding rate; second, mismatch between fast radio channel variation and information and relatively slow application adjustments.

Therefore, the application is not capable to adapt fast enough to the varying radio conditions, leading to inefficient radio resource usage and suboptimal user experience. It is proposed that coordination between the radio network and the application could be enabled to achieve more efficient use of resources and better user experience. Moreover, if application servers were deployed closer to the RAN edge, more real-time coordination could be enabled.

To be specific, in video services, clients could choose the optimum segments based on estimated bandwidth. Existing bandwidth estimation algorithms predict bandwidth with throughput in client side. However, these methods would be less sensitive to the variations in wireless network parameters, e.g. radio channel, network congestion, since there is a mismatch between millisecond-level radio variation and much slower video application adjustments. Therefore, it is necessary to optimize video delivery by allowing video and RAN mutual awareness.

As already captured in TR 38.913, "Study on Scenarios and Requirements for Next Generation Access Technologies", the RAN architecture shall allow to enable context-aware service delivery. Service-specific network optimization would be further enabled in later part of 3GPP Release 15 and Release 16.

C) Flexible MAC: grant based or grant free

With potentially explosive growth of machine type communication (MTC) applications and devices, such as sensors, meters, wearable devices, etc., highly diverse traffic profiles with trillions of wireless nodes may come to life in the fifth-generation communication system, including millions of data applications for smart phones. These devices, however, cannot be handled efficiently by current wireless communication networks, which were not designed for low latency, frequent small-data packets, and simultaneous massive accesses. Other than the existing grant-based transmission in cellular networks, grant-free transmission is proposed as a promising method to reduce transmission latency and excessive signaling overhead caused by vast small-data packet traffic, which is often UL dominant. To improve efficiency of grant-free resource allocation, grant-free resource may be shared by multiple users and may become contention based.

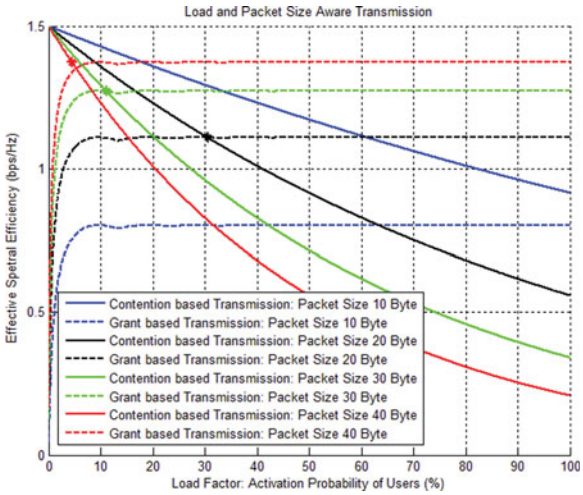

Fig. 6. Load and packet size aware transmission.

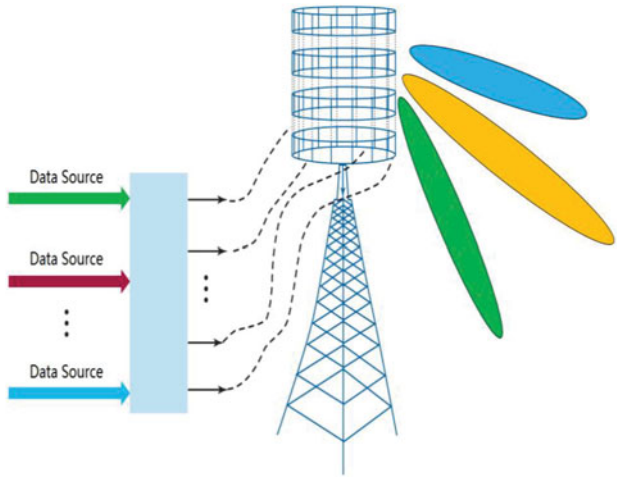

Fig. 7. The massive antenna array.

It is observed in Fig. 6 that: (1) efficiency of contention-based transmission degrades with increasing packet sizes and traffic load, while efficiency of grant-based transmission improves by increasing packet sizes and is stable over different loading; (2) shown in asterisk of the figure, for packets of 20, 30, and 40 Bytes, there is a turning load threshold that below which contention-based transmission is more efficient than grant-based transmission; (3) when the packet size is small enough, e.g. 10 Bytes, contention-based transmission is always optimal within certain load. Based on the analysis, KA messages is better suited to contention-based transmission, while the rest of the mobile Internet packet types (e.g. video or voice) need to be transmitted by grant if not taking into account latency constraints. This is because the former are of relatively small packet size and small loading, while packet size of the latter are too large to be transmitted by contention.

Currently, grant-free approach is discussed under framework of 3GPP NR. Semi-static scheduling (SPS) is adopted as a baseline approach to achieve grant free. Multiple users are allowed to share SPS. 5G is embracing more flexible MAC design for diverse services.

V. RETHINK ANTENNAS

Traditional multiple antenna transmission schemes, signaling protocol, and network structure may not be sufficient and efficient in 5G; thus, fundamental rethinking in this aspect is in need. The key considerations include, e.g. theoretical and practical deployment of massive multiple input multiple output (MIMO) systems. Massive MIMO is a multi-user MIMO technology where each BS is equipped with an array of massive number of antennas, and uses these to communicate with user terminals over the same time and frequency band as shown in Fig. 7. By coherent processing of the signals, transmit beamforming can be used to focus each signal at its desired user terminal. Besides, the more antennas that are utilized, the finer the spatial focusing can be.

On the one hand, in the research filed, hardware impairment consideration, asymptotic analysis of system performance, channel state information (CSI) estimation, and especially the proper beamforming structures need to be carefully investigated to identify the optimal digital, analog, or hybrid beamforming to best meet the requirements. On the other hand, dramatic changes may be needed in reference signals design, transmit and receive scheme design, RF path calibration, channel estimation and feedback in the aspect of standardization. Besides, the much reduced power in each RF chain may bring novel RF chain design, e.g. making use of low-power low-cost terminal-grade RF-integrated circuit and the global optimal utilization of system resources with distributed massive MIMO would be greatly facilitated via C-RAN architectures. Hence, several of the most important issues in massive MIMO will be discussed as follows.

A) Non-uniform antenna array

Targeting significant capacity enhancement in 2020, the 5G network is expected to be ultra-dense with massive antennas deployed either in a distributed or centralized manner. Theoretically, massive MIMO is expected to significantly reduce the inter-cell and intra-cell interference and hence may enhance both the SE and EE. However, to accommodate a few hundred antenna and transceiver chains all on one structure (i.e. antenna panel) in a traditional cell site manner appears to be nearly impossible, given the existing challenges and increasing difficulties of site acquisition; unless moving up to the mmW band. For massive MIMO in the more desirable, lower frequency bands, we propose to fundamentally change the future scenes of cellular network: make BS invisible, by configuring the active antenna arrays in a flexible manner on the walls of city buildings and town houses as that in Fig. 8. For example, the Chinese character "中" in the China Mobile logo ("中国移动") on buildings may actually be the BS antennas in the future.

B) SE–EE co-optimization

By implementing a large number of antennas at the BS, massive MIMO systems offer a high spatial resolution that can

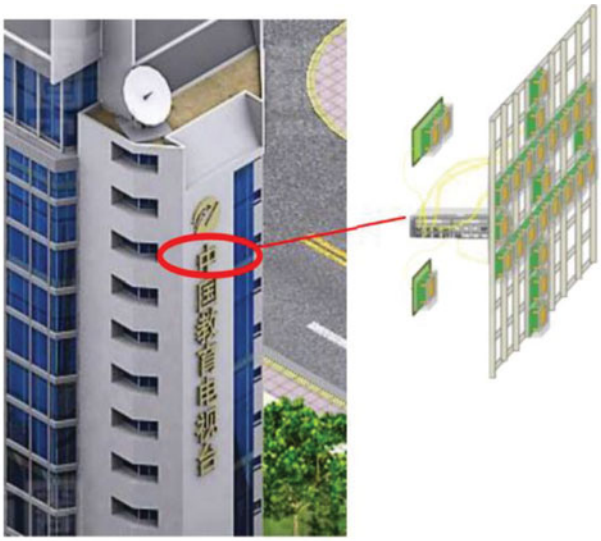

Fig. 8. One demo of non-uniform antenna array.

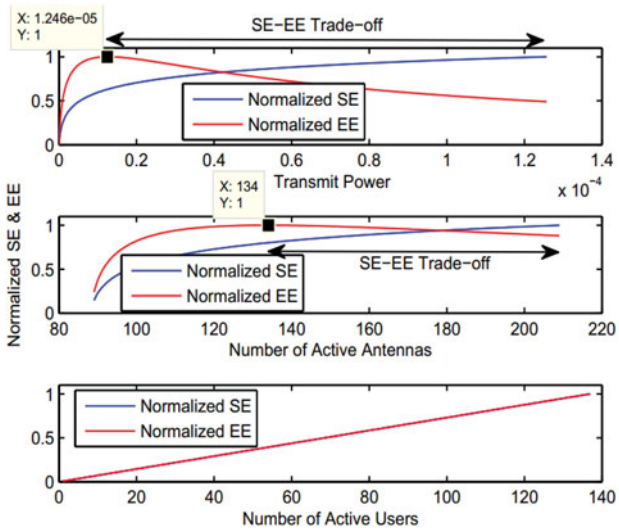

Fig. 9. The impact of p, M, and K on SE and EE.

drastically increase the spectral and/or EE of wireless systems. In contrary to the traditional MIMO systems which mainly depend on adjusting their transmit power to achieve different SE or EE, massive MIMO systems have a flexibility of changing their port number, active antenna number, active user number, and transmit power to obtain an expected SE or EE.

How to make smart use of these large amounts of antennas and users to achieve a high spectral and/or EE is a fairly new subject that is attracting substantial interest. However, maximizing one metric (EE or SE) does not mean that the other one is also maximized. In fact, the optimal EE performance often leads to low SE performance and *vice versa*.

An asymptotic analysis of SE performance has been presented for massive MIMO based on random matrix theory. In massive MIMO systems where the circuit power consumption can be comparable to or even dominates the transmit power, it would be worthwhile to investigate whether massive MIMO systems can outperform the systems with less antennas in EE.

After deriving a closed-form expression of the optimal value of transmit power (p), the number of active antennas (M) and the number of the active users (K) for global SE (or EE) maximization, we find it feasible for real-time adaptive cell planning. Besides, the EE–SE relationship would be explicit for the system operator to know whether the current cell status has maximized both SE and EE or just achieved a SE–EE trade-off.

From the simulations of different sets of $\{p, M, K\}$ (i.e. by changing the value of one parameter with other two parameters fixed with a randomly generated set of $\{P_{max}, M_{max}, K_{max} = 1 : 26e - 4, 209, 88\}$ as constraint) for maximizing EE and SE, we generate a figure for a typical occasion for them as show in Fig. 9. We could determine the optimal values of M, K, and p to maximize EE and their value for maximizing SE in this condition, and also point out how to adjust the trade-off between SE and EE.

C) Hybrid MIMO architecture

Communication over mmW frequencies will be a key feature of the next-generation (5G) cellular networks. One advantage of mmW communication is its high data rate due to the large potentially available bandwidth, which meets the high peak data rate requirements of next-generation wireless systems. Another potential advantage of mmW communication is its low latency, which is essential for many 5G applications, like V2V communications. Massive MIMO systems are also taken as one of the key architectural features of mmW communications. These antenna arrays are used to provide array gain wide area operation. Unlike traditional lower frequency MIMO systems, these large arrays combined with high cost and power consumption of the mixed analog/digital signal components make it difficult to provide each antenna with an individual RF chain, and proceed all the signal processing in the baseband. This motivates us to propose new transceiver structures and beamforming strategy, e.g. hybrid beamforming [10].

To solve this problem (i.e. to constrain the cost and power consumption), analog beamforming is one approach, which relies entirely on RF domain processing to reduce the number of RF chains. The beamforming is implemented using networks of analog phase shifters that change the relative phases of the data signals to the antennas to generate the desired directions for the UEs. While, an alternative approach to reduce the number of RF chains is hybrid analog/digital architectures. In hybrid architectures, MIMO processing is divided into the analog and digital domains to reduce the number of required transceivers, as shown in Fig. 10. Thanks to the precise digital processing, more degrees of freedom are available for the design of the hybrid beamformers compared with analog-only beamformers, allowing them to support multi-stream and multi-user transmission.

The proposed hybrid MIMO architectures for mmW communications are usually based on phased arrays

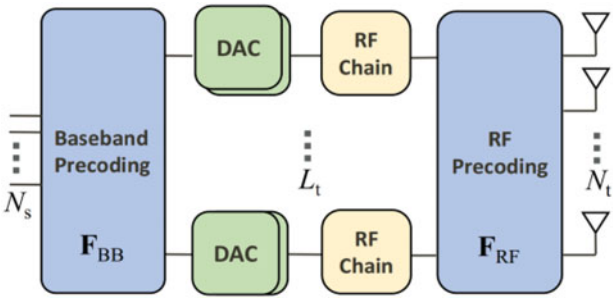

Fig. 10. Hybrid MIMO architecture for mmW communication.

(though some fairly new work is based on lens architectures as an alternative approach). Practical phased arrays use finite precision phase shifters, which may make it difficult to finely generate the beams and null space. Besides, only one low-noise amplifier (LNA) is needed in conventional MIMO receivers for each antenna, yet, mmW receivers based on phased shifters need larger numbers of LNAs to obtain the same signal-to-noise ratio at the input of the RF chains. Moreover, increasing the number of bits in the phased array leads to higher power consumption and higher complexity for the whole system. Hence, to obtain the beamforming schemes with low cost and high performance, more works should be done in this field for maturing the design of hybrid MIMO architecture in the future days.

The standardization of 5G NR MIMO is mainly focused on hybrid beamforming, which includes a unified CSI measurement and feedback framework, beam management schemes, new phase tracking reference signals, codebook design for analog and digital beamforming, etc.

VI. RETHINK SPECTRUM AND AIR INTERFACE

Since there may be many use cases emerging in 5G and beyond, it is very important for operators to deploy one network to support all use scenarios and use cases. Toward this end, it is critical to adopt one unified and flexible air interface framework to meet diverse requirements of the key 5G usage scenarios, e.g. eMBB, URLLC and mMTC. The unified framework of SDAI [14] will meet the diverse demands in 5G by reconfiguring combinations of the physical layer building blocks, including frame structure, duplex mode, waveforms and multiple access scheme, modulation and coding and spatial processing scheme. The resources at different frequency bands can be harmoniously utilized in SDAI with efficient inter-carrier coordination or joint scheduling.

A) Flexible frame structure

Frame structure is the basic DL and UL operation framework for wireless communication systems, which specifies where and when the signaling, control, and data should be transmitted. A unified frame structure concept is proposed in [15] which is capable of dealing both with broadband data services and small packet services within the same band. In order to realize SDAI, the frame structure should be flexible enough. For example, the time and frequency resources are allocated to different users with different service requirements, channel conditions, UE capabilities (multiple access support, full duplex mode, feature or smart phones), mobility, and frequency bands, etc. In different resource blocks, different air interface solutions with different multiple access schemes, transmission time interval (TTI) parameters, waveforms, and duplex mode, pilot signals can be defined. This is very challenging, since the inter-subcarrier band interference between different resource blocks needs to be carefully mitigated.

The designs on flexible frame structure have been heatedly discussed in 3GPP NR. The key features, such as scalable numerology, configurable subframe direction (DL, UL, bidirection), flexible scheduling unit (slot/mini-slot), flexible scheduling and hybrid automatic repeat request timing, flexible reference signal configurations, and efficient URLLC and eMBB service multiplexing, etc., are agreed to be supported for wide range of services requirements, deployments scenarios, and spectrum.

B) Flexible waveforms

Orthogonal frequency division multiplexing (OFDM) has been used extensively in 4G and is still considered as an important candidate waveform for 5G. However, 5G will not only continue to focus on the mobile broad-band services, but also will embrace diversified types of Internet of things services, such as MTCs. It is not adequate to only use OFDM to deal with the diversified services, higher SE, and massive connections. To meet these requirements, several new multi-carrier modulation schemes have been proposed, e.g. unified frequency multi-carrier, generalized frequency division multiplexing and filter bank multi-carrier [16, 17], and Filter-OFDM (F-OFDM) [18]. The flexible compatible framework for theses waveforms can be based on the carrier/waveform aggregation. Different waveforms located in different carriers can be aggregated in one air interface serving diverse 5G services. The waveform, sub-band bandwidth, subcarrier spacing bandwidth, filter length, and cyclic prefix length in each wave can be flexibly chosen according to the dedicated scenarios and services.

A compatible multi-carrier modulation structure with low complexity is depicted in Fig. 11, where several

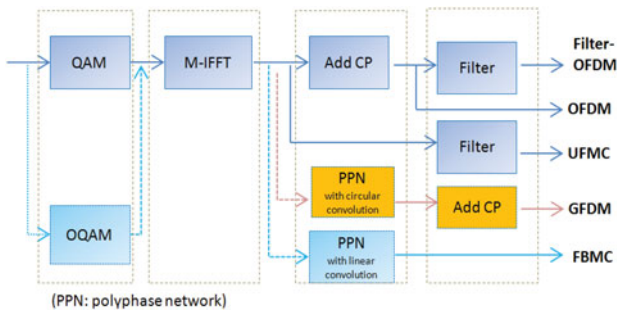

Fig. 11. A unified framework of flexible waveforms.

waveforms are generated. We can flexibly configure different waveform schemes according to various 5G scenarios on the basis of minimizing the hardware functional module.

The recent progress in 3GPP NR standardization is that cyclic prefix-OFDM (CP-OFDM) will be adopted for both DL and UL, while discrete Fourier transform-spread-OFDM (DFT-S-OFDM) will also be adopted for UL. The UE transparent waveforms are also supported in 3GPP R15, e.g. F-OFDM, wideband-OFDM (W-OFDM).

C) Flexible multiple access

Conventional Orthogonal Multiple Access (OMA) such as Frequency Division Multiple Access (FDMA), Time Division Multiple Access (TDMA), and Orthogonal Frequency Division Multiple Access (OFDMA) have been studied for years and are widely adopted in wireless communication systems from 1G to 4G. OMA schemes can conveniently support high-data-rate transmission, which capitalizes on the orthogonality and synchronization. On the other hand, the advanced multiple access technology has been envisioned as one of key enablers of 5G communication. In non-orthogonal transmission schemes [19–21], the signals from different users will be superposed into the same time and frequency resource and demodulated by advanced receiver algorithm to provide higher SE and system capability. Meanwhile, grant-free transmission will be allowed to significantly reduce signaling overhead, shorten access latency, and decrease terminal power consumption. Code-based Non-Orthogonal Multiple Access (NOMA) schemes may provide a more reliable transmission than OMA because of having more transmission chance. The multiple access techniques as introduced in literatures are summarized as below.

The above discussed advanced multiple access schemes as well as the traditional OMA scheme, e.g. OFDMA are both identified as potential candidates for 5G. Based on the diverse deployment scenarios and traffic requirements in 5G, flexible multiple access can be utilized to meet the verified demands [22]. For example, in the case of massive connections, how to accommodate more users with limited resources has become a critical problem for next-generation access network. With non-OMA schemes, e.g. Sparse Code Multiple Access (SCMA) [20]. Multi-User Shared Access (MUSA), Pattern Division Multiple Access (PDMA), or Resource Spread Multiple Access (RSMA), Bit Division Multiplexing (BDM), the same resources are shared and reused by multiple users, thus the number of connections increases. To support the traffic with low-latency requirement, non-OMA schemes help to realize grant-free multiple access, with which the latency is much lower, and the power consumption of the devices can be reduced. In other scenarios, such as DL machine-type traffic, the simple OMA schemes are better due to the device cost and implementation complexity (Table 1).

In 3GPP NR discussions, 15 NoMA schemes are proposed by different companies; but these schemes can be presented concisely in a unified framework [22]. Synchronous/scheduling-based OMA is supported for UL and DL transmissions, at least targeting for eMBB. NR targets to support UL non-OMA, in addition to the orthogonal approach. The new study item about NOMA targeting to 3GPP R16 is going to launch from the second half of year 2017.

D) Flexible duplex mode

Duplex modes have been studied as the basis of cellular networks during past several decades. Frequency division duplex (FDD) and time division duplex (TDD) have been widely used in current LTE systems. To well adapt to the traffic imbalance between UL and DL transmissions, flexible FDD and dynamic TDD are proposed [23]. To further improve the network capacity, full duplex has drawn much attention because it has the potential to maximally double the spectral efficiency. The BSs with flexible duplex are able to select the duplex mode for each frequency band, either to transmit or receive in certain time duration, or to transmit and receive simultaneously.

Frame structure is generally the basic UL and DL operation framework for wireless communication systems, which specifies where and when the signaling and data should be transmitted. The TDD frame structure consists of both UL and DL subframes on the same frequency but duplexed in

Table 1. Summary of multiple access techniques [22].

	BDM	MUSA	SPC-NOMA	PDMA	RSMA	SCMA
Scenario	DL eMBB	UL MMC, DL eMBB	eMBB, MMC, URC	eMBB, MMC, URC	UL MMC/UL URC	eMBB, MMC, URC
Multiplexing domain	Code/power	Code/power	Power	Code/power/ spatial	Code/power	Code/power
Transmitter overloading	High	High	Medium	High	High	High
Transmitter spreading	No	Yes	No	Yes	Yes	Yes
Transmitter multi-dimension constellation	No	No	No	No	No	Yes
Receiver	MMSE/SIC	SIC	SIC	SIC/MPA	SIC	MPA/SIC
Receiver complexity	Low (SSD), medium (MSD)	Medium	Medium	Medium	Medium	Medium

time domain. A guard gap is needed between the UL and DL subframes. The UL and DL frames in FDD are transmitted simultaneously on different frequency bands. These characteristics of TDD and FDD almost eliminate the necessity of joint design of UL and DL frame structure. For flexible duplex transmission, both DL and UL transmission should be supported in each time and frequency resource. To apply flexible duplex technique to the current networks early, one focus lies in leveraging full-duplex capabilities at infrastructure nodes to support half-duplex UEs, since full duplex UE still seems impractical due to complexity and cost.

VII. RETHINK PROTOCOL STACK

For the massive data scenario and the deployment of dense nodes in 5G, multi-RAT for PHY and big data computing capability based on cloud platforms are introduced. Nevertheless, the traditional LTE protocol stack is unable to complete optimal configuration for air interface resources and fails to provide UE with specific services to meet the QoS requirements. It is necessary to rethink air interface protocol stack for 5G.

Multi-level Centralized and Distributed (MCD) air interface protocol stack for 5G was proposed. In this proposal, "cell" and "UE" are managed separately. "Cell" is an element of radio resource management and provides appropriate radio resources for UE to fulfill the requirement of air interface, and also implements flexible resource control. Radio resources of a cell are divided into two types: inter-cell and intra-cell. According to the characteristics of radio resources, a combination of fast and slow management increases utilization of resources. The context and data of UE are separated, and each of them was centrally managed to achieve unity of computing capability, which is a better fit for cloud platform.

Although the framework of traditional LTE protocol stack is defined in details, the signaling in protocol stack is complex. The LTE protocol stack architecture is unable to support high-density 5G network, massive users, and various kinds of services in 5G. We need to rethink the protocol stack architecture. The protocol stack architecture should be "user-centric", and provide flexible air interface and reduce the frequency of RRC signaling transmission. Meanwhile, the protocol stack architecture should take full advantage of "cloud" with enormous computing capability. Considering the high density of users and cells, with big data, the protocol stack architecture should implement the optimized configuration over air interface resources, e.g. frequency resources, time resources, and space resources.

For traditional LTE/LTE-A, the fundamental element of communication network is "cell", which manages the radio resources and the users connected to it. In traditional LTE protocol, which is shown in Fig. 12, the UE context can only be established and managed within a specific cell. In the case of carrier aggregation (CA), the UE context is established within the PCell rather than secondary cells (SCells). Furthermore, the SCells only provide channels

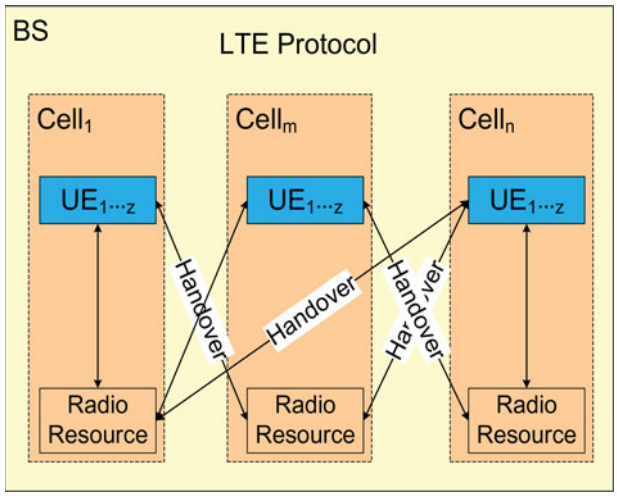

Fig. 12. LTE protocol stack.

for data transmission/receiving. During handover procedure between PCell and SCell, the signaling is complex and latency is a scale of several seconds or even minutes. However, the serving cells for UE are unchanged. What is more, for some technologies, the signaling transmission is semi-static, e.g. Inter Cell Interference Coordination.

In 5G network, user-centric network (UCN) is introduced to solve the problem of explosive growth of data traffic and increasing density of BSs. The signaling transmission in UCN is in the way of control C/U decoupling. According to the quality of channels, the network should provide the corresponding radio services in order to maintain the connection of control plane and user plane and transmission of signaling and data. To improve the quality of air interface, many new technologies will be introduced to 5G network, e.g. full duplex, hybrid PHY, and so on. Nevertheless, those new technologies bring many challenges to 5G, which means the network should provide corresponding services to UE to meet the requirements for data rate and channel quality in every TTI. In order to satisfy the requirements, the coordination among cells should keep real time in each TTI, which greatly increases the difficulty of processing on the network side.

It is necessary to rethink the air interface protocol stack for the requirements of 5G and the status of traditional networks. The traditional network, which is characterized as cell centric, has been proved to be a simple and practical method of radio resource management, and adapts to the framework of cellular network [24]. The air interface protocol stack for 5G should inherit the advantages of traditional networks, and be rethought to meet the requirement of 5G network and coexist with traditional network.

The difference between MCD protocol stack and traditional protocol stack is showed in Figs 12 and 13. In traditional protocol stack, the management unit of signaling and UE context is cell, e.g. the scope of Cell Radio Network Temporary Identifier for each cell is 0–65 535 [25]. DRB and SRB of UE are both allocated and managed in the scope cells, as well as the process of DRB and SRB mapping to E-UTRAN

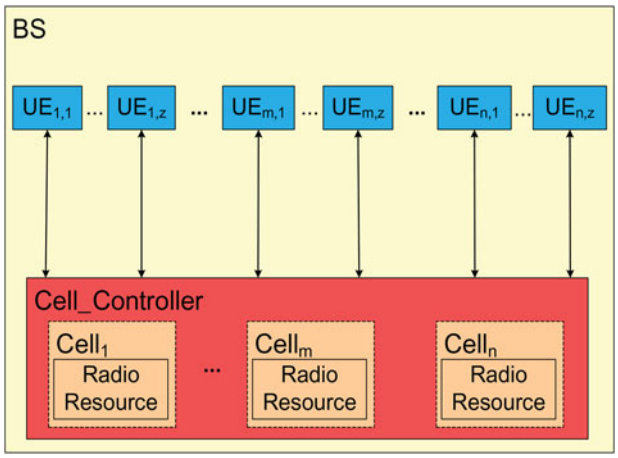

Fig. 13. Rethink protocol stack.

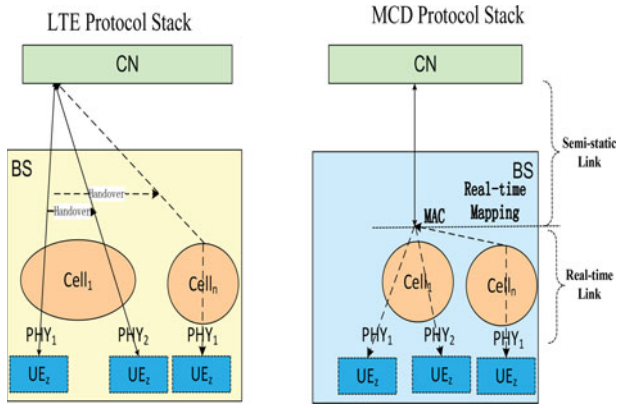

Fig. 14. Mobility of UEs.

Radio Access Bearer (E-RAB) [24]. In the case of CA, UE can use the resource of more than one cell. CA cannot assist LTE network to solve the problem of 5G [25]. In summary, on the one hand, with "cell" as the key label, LTE protocol stack simplifies the process of radio resource management for the network. On the other hand, such protocol stack increases the complexity of management and lead to a high latency when UE moves and is hard to fit the needs of 5G.

In MCD protocol stack, "UE" is also a basic element as well as "cell". On the one hand, as an element of protocol stack, UE responds for the management of all the information itself, including UE context, the mapping process between DRB and SRB to E-RAB, channel quality and the dedicated radio resource allocated to UE, and so on. On the other hand, as another element of protocol, cell manages all radio resources that are not allocated to any users. As shown in Fig. 13 the Cell_Controller module manages cells, which allocate their own radio resource based on the allocation result of Cell_Controller. According to the specific requirements of each UE, Cell_Controller allocates corresponding cells to it. Those cells with available radio resources can fulfill the demand of UE. The radio resource which is allocated to UE becomes a specific attribute of UE, and UE implements the management of radio resource. UE releases radio resource to cell when transmission process ends. Resource allocation and release are just changes of UE's attribute as a setup operation, which works in the same way as UE context modification. Such a setup operation avoids the changes of DRB and logical channels when radio resource changes. From the macroscopic perspective, handover procedure is replaced by modification of UE radio resource attribute, which implements via faster radio resource deployment of air interface during UE moves across cells. Our proposal, in which the DRB and logical channels remain unchanged and the radio resource deployment changes fast, fulfills the need of 5G network, such as ultra-dense coverage, smaller delay, and high reliability, and so on.

To meet the requirements of 5G network which characterize high-density cells, vast amounts of user and hybrid PHY, the MCD protocol stack of 5G air interface with "UE" and "cell" as basic elements provides a pattern for link controlling. As shown in Fig. 14, this pattern with slow combination of semi-static link for UEs module and real-time channel mapping for UEs module aims to realize o ms-lantency handover.

The semi-static link for UEs module focus on logical channels, DRB/SRB and E-RAB link, all of those links indicate the specific type of service (ToS) of UE. To implement unbundling ToS with a specific PHY mode in hybrid PHY, those links work in the way of semi-static control. When the protocol runs under C-RAN, those links only work with the modes of establishment, reconfiguration, and deletion. When an UE moves to the coverage of another C-RAN, handover procedure is implemented by redefining handover signaling. In this way, semi-static pattern realize the o ms-lantency handover in the scenario of high-density cells.

In real-time channel mapping for UEs module, the mapping of logical channels to transport channels and transport channels to physical channels are in real-time pattern [26]. A UE provides parameters to MAC such as quality of channel, buffer occupy, request for PHY, and the characters of allocated radio resource. According to radio resource of all the available cells and the parameters received, MAC configures cell and its radio resource as attributes to the UE. In that pattern, logical channels match the appropriate cell at first, then map to transport channels, and transport channels map to physical channels in that cell. With real-time resources modification rather than handover procedure, the UE can receive data from different cells with o ms interruption time.

With the combination of semi-static link for UEs module in low speed and real-time channel mapping for UEs module in fast speed, the pattern proposed implements MCD control and radio resource allocation. As shown in Fig. 14 in traditional LTE protocol stack architecture, handover procedure is necessary when the UE moves across cells. Nevertheless, in 5G protocol stack architecture characterized as MCD, the real-time mapping replaces handover procedure. Furthermore, handover of semi-static link for UEs module takes place only when UEs move to another C-RAN. Flexible control of air interface will be implemented in 5G protocol stack architecture by decreasing the frequency and

complexity of handover signaling procedure as well as the latency and increasing the reliability at the same time.

The MCD design logic is playing important role during the standardization of 5G NR. Based on the discussions of 3GPP RAN2 study items and work items, the MCD design logic has been embodied by the concept of two-level mobility, i.e. RRC-configured level and MAC-assisted level. Particularly, MAC-assisted mobility requires the cell definition clarification and the protocol stack enhancement. In RAN3, CU/DU functional split, for which non-real-time functions are allocated on CU, while real-time functions are allocated on DU, is exactly a reflection of the MCD design logic.

VIII. RETHINK FH

With data bandwidth expected to continue to grow exponentially and as the mobile wireless industry moves onwards from 4G networks toward 5G, it is becoming clear that the existing FH infrastructure using the common public radio interface (CPRI) protocol is not going to scale in the existing topologies in use, let alone address future network topologies. In [5, 28, 29], the authors proposed to redefine the CPRI and brought forward a new concept called next-generation fronthaul interface (NGFI). NGFI possesses the following desirable features.

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

The key way to achieve NGFI is function re-split between the BBUs and the remote radio units (RRUs). Traditionally, all the baseband functions, including the PHY, MAC, and PDCP are processed on the BBU side, while the RRU mainly deals with the radio-related functions. The signal transmitted by CPRI is the high-bandwidth I/Q sampling signal. From the effective information perspective, any data between the baseband protocol stacks (e.g. between MAC and PHY) could be transported. The basic idea of function splitting is to move partial baseband functions to the RRU to reduce the bandwidth without losing any information.

There have been some related studies in literature on this topic. To achieve NGFI in general, the function splitting should decouple the bandwidth from the antennas, which can be achieved by moving antenna-related functions (DL antenna mapping, FFT, channel estimation, equalization, etc.) to the RRH. In this case, it was then shown in [29] that the FH bandwidth of an LTE carrier may decrease to the order of 100 Mb/s no matter how many antennas are used. In addition, it is suggested that the UE processing functions should be decoupled from cell-processing functions. In this way, the FH bandwidth will be lowered and more importantly, load-dependent. The load-dependent feature gives an opportunity to exploit the statistical multiplexing gain when it comes to FH transport network design for C-RAN deployment. Thanks to statistical multiplexing, the bandwidth needed for transport of a number of FH links in C-RAN can be reduced greatly, subsequently decreasing the cost.

Support of collaborative technologies is another key factor for the design of function splitting. Coordinated multi-point (CoMP) has been viewed as one of the key technologies in 4G and 5G to mitigate the interference. CoMP can be divided into two classes: MAC layer coordination and physical layer coordination. For example, collaborative schedule is one of the MAC layer-coordinated mechanisms. Joint reception (JR) and joint transmission (JT) are the physical layer-coordinated technologies. In [30], it was found that the performance gain of JR/JT decreases significantly as the number of antennas increases. Moreover, in [31] the authors found through field trial data that MAC-level collaborative technologies can bring comparable performance gains with lower complexity, easier implementation, and fewer constraints. Based on the observations, it is suggested that the function splitting for NGFI does not have to support PHY layer coordination technology. It is enough to achieve considerable performance gain by supporting MAC layer-coordinated technologies.

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

The main challenges for the Ethernet as a FH solution lie on the high timing and synchronization requirements imposed by the NGFI interface. Although the exact NGFI has so far not been specified, it is possible that NGFI may keep some requirements of CPRI, such as synchronization requirements. The allowable RF error for a CPRI link is ± 2 ppb and the timing alignment error shall not exceed 65 ns in order to support MIMO and transmission diversity. In order to meet the timing requirements, both the BBU and the RRUs should be perfectly synchronized, which therefore requires a very accurate clock distribution mechanism. Potential solutions may include any combination of Global Positioning System (GPS), IEEE 1588, and synchronous Ethernet. Finally, the transport protocols on top of the Ethernet such as Multi-Protocol Label Switching and Packet Transport Network that establish transport paths for FH traffic

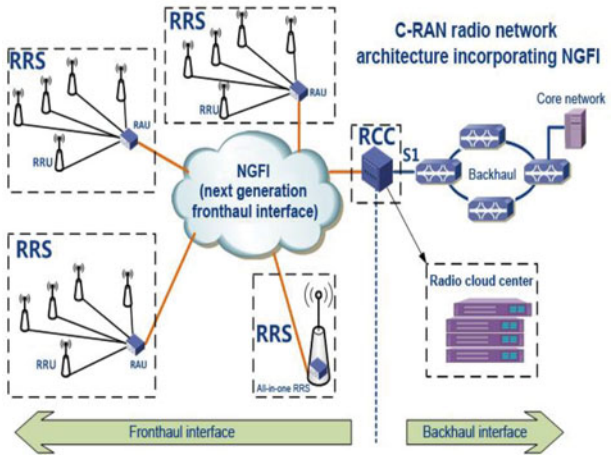

Fig. 15. NGFI-based C-RAN architecture. RRS, Remote Radio System; RAU, Radio Aggregation Unit.

need to be defined. Furthermore, SDN technology should also be integrated to further improve the transport network efficiency and flexibility.

The C-RAN architecture also evolves as traditional FH interfaces change to NGFI. As shown in Fig. 15, the evolved C-RAN consists of three parts:

(1) Radio Aggregation Unit (RAU): With function split, the moved partial BB functions form a new entity which is called the RAU. RAU is a logical concept and its realization depends on implementation solutions. For example, RAU could be integrated into the RRH to form a new type of RRH. Alternatively, it could also be an independent hardware entity.
(2) RSS: A RRS consists of an RAU and multiple RRHs. It is expected that collaboration could happen among different RRHs via the RAU within the same area coverage of a RRS. There could be multiple RRS in a C-RAN network.
(3) RCC: The remaining BB functions together with high-layer functionalities constitute a RCC. RCC is the place where all the processing resources are pooled into a cloud with virtualization technology.

Since the proposal of NGFI concept, there has been a consensus and great interest from the industry. Currently several organizations are dealing with NGFI standardization. In IEEE, a 1914 NGFI working group has been founded in 2016, studying the architecture and requirement development from transport perspective. Another group, 802.1 CM is addressing the transport solutions to guarantee the real-time requirements by NGFI transport. Meantime, 3GPP is studying various function split options, which is critical to NGFI implementation.

IX. CONCLUSIONS

Green, soft, and super-fast have been recognized as key features for future 5G wireless systems. This paper gave an overview of such vision and potential solutions. The 5G network design considerations were elaborated, with seven fundamental rethinking on the Shannon theory, Ring and Young, signaling and control, antennas, spectrum and air interface, protocol stack and FH.

ACKNOWLEDGEMENTS

This paper is supported by the National High Technology Research and Development Program (863 Program) of China No. 2015AA01A702. The authors would like to thank the team members in the Green Communication Research Center of China Mobile Research Institute, particularly Sen Wang, Gang Li, Yami Chen, Guozhen Xu, Junshuai Sun, Jinri Huang, Qi Sun, and Shuangfeng Han.

REFERENCES

[1] I, C.-L.; Rowell, C.; Han, S.; Xu, Z.; Li, G.; Pan, Z.: Towards green & soft: A 5G perspective. IEEE Commun. Mag., 52 (2), (2014), 66–73.

[2] Chen, S.; Zhao, J.: The requirements, challenges, and technologies for 5G of terrestrial mobile telecommunication. IEEE Commun. Mag., 52 (5), (2014), 36–43.

[3] GSMA Intelligence: Understanding 5G: Perspectives on future technological advancements in mobile, white paper, 2014.

[4] IMT-2020 (5G) promotion group (PG): 5G vision and requirements, May 2014. Available: http://www.imt-2020.org.cn/zh/documents.

[5] CMRI: C-RAN: the road towards green RAN, October 2011. Available: labs.chinamobile.com/cran.

[6] Introduction White Paper: Network functions virtualization, October 22–24, 2012 at the "SDN and OpenFlow World congress", Darmstadt-Germany.

[7] ONF White Paper: Software-defined networking: the new norm for networks, April 2012.

[8] Xu, Z.; Pan, Z.; I, C.-L.: Fundamental properties of the EE-SE relationship, in IEEE WCNC, 2014, 1115–1120.

[9] Li, G.Y.; *et al.* Energy-efficient wireless communications: tutorial, survey, and open issues. IEEE Wireless Commun., 18 (6), (2011), 28–35.

[10] Han, S.; I, C.-L.; Xu, Z.; Rowell, C.: Large scale antenna systems with hybrid analog and digital beamforming for millimeter wave 5G. IEEE Commun. Mag., 53 (1), (2015), 186–194.

[11] Bhushan, N. *et al.*: Network densification: the dominant theme for wireless evolution into 5G. IEEE Commun. Mag., 52 (2), (2014), 82–89.

[12] NGMN Alliance: NGMN KPIs and deployment scenarios for consideration for IMT2020 v1.0, December, 2015.

[13] Chen, Y.; Li, G.; Pan, Z.; I, C.-L.: Small data optimized radio access network signaling/control design, in ICC WS, 2014, 49–54.

[14] Sun, Q.; Han, S.; I, C.-L.; Pan, Z.: Software defined air interface A framework of 5G air interface, in IEEE WCNC, 2015, 6–11.

[15] Wunder, G. *et al.*: 5GNOW: Intermediate frame structure and transceiver concepts, in IEEE GC Workshop, 2014, 565–570.

[16] Wunder, G. *et al.*: 5GNOW: non-orthogonal, asynchronous waveforms for future mobile applications. IEEE Commun. Mag., 52 (2), (2014), 97–105.

[17] PHYDAYS: FBMC physical layer: a primer, June 2010.

[18] Abdoli, J.; Jia, M.; Ma, J.: Filtered OFDM: A new waveform for future wireless systems, in IEEE Signal Processing Advances in Wireless Communication (SPAWC), July 2015, 66–70.

[19] Dai, L.; Wang, B.; Yuan, Y.; Han, S.; I, C.-L.; Wang, Z.: Non-Orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends. IEEE Commun. Mag., 53 (9), (2015), 74–81.

[20] Nikopour, H.; Baligh, H.: Sparse code multiple access, in Proc. IEEE PIMRC, 2013, September 2013, 332–336.

[21] Saito, Y.; Kishiyama, Y.; Benjebbour, A.; Nakamura, T.; Li, A.; Higuchi, K.: Non-orthogonal multiple access (NOMA) for future radio access, in Proc. IEEE VTC Spring 2013 June 2013, 1–5.

[22] R1-162870: On unified framework for multiple access schemes, CMCC.

[23] DUPLO Deliverable D4.1.1: Performance of full duplex systems, January 31, 2014.

[24] TS 36.401: E-UTRAN Architecture description (Release 11), Version 12.2.0, March 2015.

[25] TS 36.410: E-UTRAN S1 General aspects and principles (Release 11), Version 12.1.0, December 2014.

[26] TS 36.413: E-UTRAN S1 Application Protocol (Release 11), Version 12.6.0, June 2015.

[27] TS 36.300: E-UTRAN Overall description (Release 11), Version 12.6.0, June 2015.

[28] TS 36.211: E-UTRAN Physical Channels and Modulation (Release 11), Version 12.6.0, June 2015.

[29] I, C.-L.; Huang, J.; Yuan, Y.; Ma, S.; Dan, R.; Cui, C.: Rethink fronthaul for soft RAN. IEEE Commun. Mag., 53 (9), (2015), 82–88.

[30] I, C.-L.; Huang, J.; Yuan, Y.; Ma, S.; Dan, R.: NGFI, the xHaul, in IEEE Globecom, December 2015, 1–6.

[31] Davydov, A.; Morozov, G.; Bolotin, I.; Papathanassiou, A.: Evaluation of joint transmission CoMP in C-RAN based LTE-A HetNets with large coordination areas, in IEEE globecom workshops, December 2013, 801–806.

CHIH-LIN I received her Ph.D. degree in electrical engineering from Stanford University. She has been working at multiple world-class companies and research institutes leading the R&D, including AT&T Bell Labs; Director of AT&T HQ, Director of ITRI Taiwan, and VPGD of ASTRI Hong Kong. She received the IEEE Trans. COM Stephen Rice Best Paper Award, is a winner of the CCCP National 1000 Talent Program, and has won the 2015 Industrial Innovation Award of IEEE Communication Society for Leadership and Innovation in Next-Generation Cellular Wireless Networks. In 2011, she joined China Mobile as its Chief Scientist of wireless technologies, established the Green Communications Research Center, and launched the 5G Key Technologies R&D. She is spearheading major initiatives including 5G, C-RAN, high energy efficiency system architectures, technologies and devices; and green energy. She was an Area Editor of IEEE/ACM Trans. NET, an elected Board Member of IEEE ComSoc, Chair of the ComSoc Meetings and Conferences Board, and Founding Chair of the IEEE WCNC Steering Committee. She was a Professor at NCTU, an Adjunct Professor at NTU, and an Adjunct Professor at BUPT. She is the Chair of FuTURE 5G SIG, an Executive Board Member of GreenTouch, a Network Operator Council Founding Member of ETSI NFV, a Steering Board Member and Vice Chair of WWRF, a Steering Committee member and the Publication Chair of IEEE 5G Initiative, a member of IEEE ComSoc SDB, SPC, and CSCN-SC, and a Scientific Advisory Board Member of Singapore NRF. Her current research interests center around "Green, Soft, and Open".