Cloud Radio Access Network: Virtualizing Wireless Access for Dense Heterogeneous Systems

Osvaldo Simeone, Andreas Maeder, Mugen Peng, Onur Sahin, and Wei Yu

Abstract: **Cloud radio access network (C-RAN) refers to the virtualization of base station functionalities by means of cloud computing. This results in a novel cellular architecture in which low-cost wireless access points, known as radio units or remote radio heads, are centrally managed by a reconfigurable centralized "cloud", or central, unit. C-RAN allows operators to reduce the capital and operating expenses needed to deploy and maintain dense heterogeneous networks. This critical advantage, along with spectral efficiency, statistical multiplexing and load balancing gains, make C-RAN well positioned to be one of the key technologies in the development of 5G systems. In this paper, a succinct overview is presented regarding the state of the art on the research on C-RAN with emphasis on fronthaul compression, baseband processing, medium access control, resource allocation, system-level considerations and standardization efforts.**

Index Terms: **Backhaul, cloud radio access networks, common public radio interface (CPRI), cloud radio access network (C-RAN), 5G, fronthaul, radio resource management.**

I. INTRODUCTION

CLOUD radio access network (C-RAN) refers to the virtualization of base station functionalities by means of cloud computing. In a C-RAN, the baseband and higher-layers operations of the base stations are implemented on centralized, typically general-purpose, processors, rather than on the local hardware of the wireless access nodes. The access points hence retain only radio functionalities and need not implement the protocol stack of full-fledged base stations. This results in a novel cellular architecture in which low-cost wireless access nodes, known as radio units (RUs) or remote radio heads (RRHs), are centrally managed by a reconfigurable centralized "cloud", or central, unit (CU). At a high level, the C-RAN concept can be seen as an instance of network function virtualization (NFV) techniques and hence as the RAN counterpart of the separation of control and data planes proposed for the core network in software-defined networking (see, e.g., [1]).

Manuscript received September 15, 2014.

The work of O. Simeone has been partly supported by U.S. NSF under grant CCF-1525629. The work of W. Yu has been supported in part by Natural Sciences and Engineering Research Council (NSERC) of Canada and in part by Huawei Technologies Canada Co., Ltd.

O. Simeone is with New Jersey Institute of Technology, NJ, USA, email: osvaldo.simeone@njit.edu.

A. Maeder is with NOKIA Networks, Munich, Germany, email: andreas.maeder@nokia.com

M. Peng is with Beijing University of Posts and Telecommunications, China, email: pmg@bupt.edu.cn.

O. Sahin is with InterDigital, UK, email: onur.sahin@interdigital.com.

W. Yu is with the University of Toronto, Canada, email: weiyu@comm.utoronto.ca.

Digital object identifier 10.1109/JCN.2016.000023

Referring to [2] for a discussion of the origin and evolution of the C-RAN concept, we observe here that this novel architecture has the following key advantages:

- It reduces the cost for the deployment of dense heterogeneous networks, owing to the possibility to substitute full-fledged base stations with RUs having reduced space and energy requirements;
- It enables the flexible allocation of radio and computing resources across all the connected RUs managed by the same CU, hence reaping statistical multiplexing gains due to load balancing;
- It facilitates the implementation of coordinated and cooperative transmission/ reception strategies, such as enhanced inter-cell interference coordination (eICIC) and coordinated multi-point transmission (CoMP) in long term evolution advanced (LTE-A), across the RUs connected to the same CU, thus boosting the spectral efficiency;
- It simplifies network upgrades and maintenance due to the centralization of RAN functionalities.

In a C-RAN, as mentioned, the RUs implement only radio functionalities, including transmission/reception, filtering, amplification, down- and up-conversion and possibly analog-to-digital conversion (ADC) and digital-to-analog conversion (DAC). Therefore, for the downlink, each RU needs to receive from the CU either directly the analog radio signal, possibly at an intermediate frequency, that it is to transmit on the radio interface, or a digitized version of the corresponding baseband samples. In a similar fashion, in the uplink, the RUs are required to convey their respective received signals, either in analog format or in the form of digitized baseband samples, to the CU for processing. We refer to Fig. 1 for an illustration. The RU-CU bidirectional links that carry such information are referred to as *fronthaul* links, in contrast to the backhaul links connecting the CU to the core network. The analog transport solution is typically implemented on fronthaul links by means of radio-over-fiber (see, e.g., [3]), but techniques based on copper local area network (LAN) cables are also available [4]. Instead, the digital transmission of baseband, or in-phase quadrature (IQ), samples is currently carried out by following the common public radio interface (CPRI) standard [5], which conventionally also requires fiber optic fronthaul links. The digital approach appears to be favored due to the traditional advantages of digital solutions, including resilience to noise and hardware impairments and flexibility in the transport options (see, e.g., [2]).

The main roadblock to the realization of the mentioned promises of C-RANs hinges on the effective integration of the wireless interface provided by the RUs with the fronthaul transport network. In fact, the inherent restriction on bandwidth and latency of the fronthaul links may limit the effectiveness of

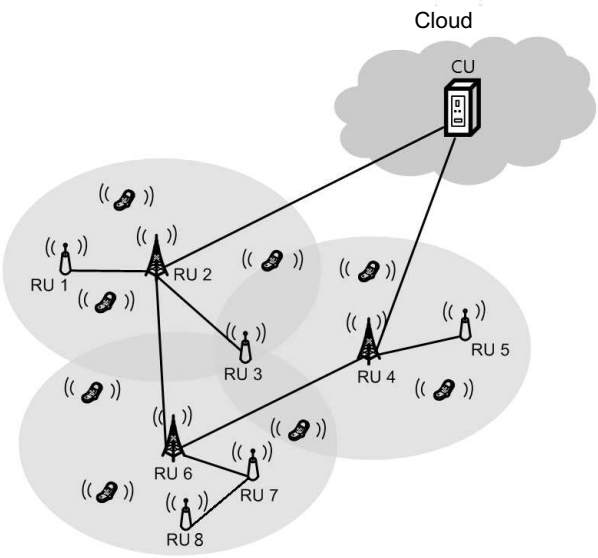


Fig. 1. Illustration of a C-RAN with a general multi-hop fronthaul network.

cloud processing. As an example, the latency induced by two-way fronthaul communication may prevent the use of standard closed-loop error recovery techniques. These problems may be alleviated by a more flexible separation of functionalities between RUs and CU whereby parts of the baseband processing, such as fast Fourier transform/inverse fast Fourier transform (FFT/IFFT), demapping and synchronization, and possibly of higher layers, such as error detection, are carried out at the RU [6], [7].

In this paper, we provide a brief overview of the state of the art on the research on C-RAN with emphasis on fronthaul compression, baseband processing, medium access control, resource allocation, system-level considerations and standardization efforts. We start in Section II with a discussion of typical baseband models used in the analysis of C-RAN systems. Then, in Section III, solutions for fronthaul transport and compression are reviewed. This is followed in Section IV by a review of relevant baseband processing techniques for C-RAN, along with the corresponding information theoretic analysis. Section V and Section VI cover design issues pertaining to higher layers, namely medium access layer and radio resource management, respectively. Section VII elaborates on architectural considerations, and Section VIII provides a short discussion on standardization efforts. Finally, Section IX closes the paper with some concluding remarks.

II. C-RAN SIGNAL MODELS

In order to introduce some of the main definitions, we start with a brief discussion of basic C-RAN signal models that are typically used in the analysis of the physical layer of C-RAN and that will be often referred to in the paper.

A. Uplink

In a C-RAN, the RUs are partitioned into clusters, such that all RUs within a cluster are managed by a single CU. Within the area covered by a given cluster, there are N_U multi-antenna user equipments (UEs) and N_R multi-antenna RUs. In the uplink, the UEs transmit wirelessly to the RUs. The fronthaul network connecting the RUs to the CU may have a single-hop topology, in which all RUs are directly connected to the CU, or, more generally, a multi-hop topology, as illustrated in Fig. 1. An example of a single-hop C-RAN is the network shown in Fig. 1 when restricted to RU 2 and RU 4.

Focusing for brevity on flat-fading channels, the discrete-time complex baseband, or IQ, signal $\mathbf{y}_i^{\rm ul}$ received by the ith RU at any given time sample can be written using the standard linear model

$$\mathbf{y}_i^{\rm ul} = \mathbf{H}_i^{\rm ul}\mathbf{x}^{\rm ul} + \mathbf{z}_i^{\rm ul} \tag{1}$$

where $\mathbf{H}_i^{\rm ul}$ represents the channel matrix from all the UEs in the cluster toward the ith RU; $\mathbf{x}^{\rm ul}$ is the vector of IQ samples from the signals transmitted by all the UEs in the cluster; and $\mathbf{z}_i^{\rm ul}$ models thermal noise and the interference arising from the other clusters. Note that in (1), and in the following, we do not denote explicitly the dependence of the signals on the sample index in order to simplify the notation. We will provide further details on the system model in Section IV.

We observe that the received signal is typically oversampled at the RUs (see Section III) and that the signal model (1) can generally account also for oversampling in the simple case under discussion of flat-fading channels. Moreover, different assumptions can be made regarding the time variability of the channel matrices depending on mobility and transmission parameters.

In the single-hop topology, each RU i is connected to the CU via a fronthaul link of capacity C_i bits/s/Hz. The fronthaul capacity is normalized to the bandwidth of the uplink channel. This implies that for any uplink coding block of n symbols, nC_i bits can be transmitted on the ith fronthaul link. In a multi-hop topology, an RU may communicate to the CU over a cascade of finite-capacity links.

B. Downlink

In the downlink, similar to the uplink, assuming flat-fading channels, each UE k in the cluster under study receives a discrete-time baseband signal given as

$$\mathbf{y}_k^{\rm dl} = \mathbf{H}_k^{\rm dl}\mathbf{x}^{\rm dl} + \mathbf{z}_k^{\rm dl} \tag{2}$$

where $\mathbf{x}^{\rm dl}$ is the aggregate baseband signal vector sample transmitted by all the RUs in the cluster; the additive noise $\mathbf{z}_k^{\rm dl}$ accounts for thermal noise and interference from the other clusters; and the matrix $\mathbf{H}_k^{\rm dl}$ denotes the channel response matrix from all the RUs in the cluster toward UE k. The fronthaul network can also be modelled in the same fashion as for the uplink. Further discussion can be found in Section IV.

III. FRONTHAUL COMPRESSION

In this section, we provide an overview of the state of the art on the problem of transporting digitized IQ baseband signals on the fronthaul links. We first review the basics of the CPRI standard in subsection III-A. Then, having identified the limitations of the scalar quantization approach specified by CPRI, subsection III-B reviews techniques that have been proposed to

reduce the bit rate of CPRI by means of compression as applied separately on each fronthaul link, i.e., via *point-to-point* compression. Finally, in subsection III-C, advanced solutions inspired by network information theory are discussed that adapt the compression strategy to the network and channel conditions by means of signal processing across multiple fronthaul links.

A. Scalar Quantization: CPRI

CPRI specification was issued by a consortium of radio equipment manufacturers with the aim of standardizing the communication interface between CU and RUs[1] on the fronthaul network. CPRI prescribes, on the one hand, the use of sampling and scalar quantization for the digitization of the baseband signals, and, on the other, a constant bit rate serial interface for the transmission of the resulting bit rate. Note that the baseband signals are either obtained from downconversion for the uplink or produced by the CU after baseband processing (see next section) for the downlink. The CPRI interface specifies a frame structure that is designed to carry user-plane data, namely the quantized IQ samples, along with the control and management plane, for, e.g., error detection and correction, and the synchronization plane data. It supports 3rd generation partnership project (3GPP) global system for mobile communications (GSM)/enhanced data rates for GSM evolution (EDGE), 3GPP universal terrestrial radio access (UTRA) and LTE, and allows for star, chain, tree, ring and multihop fronthaul topologies. CPRI signals are defined at different bit rates up to 9.8 Gbps and are constrained by strict requirements in terms of probability of error (10^{-12}), timing accuracy (0.002 ppm) and delay (5 μs excluding propagation).

The line rates produced by CPRI are proportional to the bandwidth of the signal to be digitized, to the number of receive antennas and to the number of bits per sample, where the number of bits per I or Q sample is in the range 8–20 bits per sample for LTE in both the uplink and the downlink. Accordingly, the bit rate required for LTE base stations that serves multiple cell sectors with carrier aggregation and multiple antennas easily exceeds the maximum CPRI rate of 9.8 Gbps and hence the capacity of standard fiber optic links (see, e.g., [8]). More discussion on CPRI can be found in Section VIII.

B. Point-to-Point Compression

As discussed, the basic approach prescribed by CPRI, which is based on sampling and scalar quantization, is bound to produce bit rates that are difficult to accommodate within the available fronthaul capacities — most notably for small cells with wireless fronthauling and for larger cells with optical fronthaul links in the presence of carrier aggregation and large-array MIMO transceivers. This has motivated the design of strategies that reduce the bit rate of the CPRI data stream while limiting the distortion incurred on the quantized signal. Here we provide an overview of these schemes by differentiating between techniques that adhere to the standard C-RAN implementation with full migration of baseband processing at the RU and solutions that explore different functional splits between RU and CU.

[1]The terminology used in CPRI is radio equipment control (REC) and Radio Equipment (RE), respectively.

B.1 Compressed CPRI

In the first class, we have techniques that reduce the CPRI fronthaul rate by means of compression. The so called compressed CPRI techniques are based on a number of principles, which are briefly discussed in the following.

1) *Filtering and downsampling* [9], [10]: As per the CPRI standard, the time-domain signal is oversampled. For instance, for a 10 MHz LTE signal a sampling frequency of 15.36 MHz is adopted. Therefore, a low-pass filter followed by downsampling can be applied to the signal without affecting the information content.

2) *Per-block scaling* [9], [10]: In order to overcome the limitations due to the large peak-to-peak variations of the time-domain signal, per-block scaling can be performed. Accordingly, the signal is divided into subblocks of small size (e.g., 32 samples in [9]) and rescaling the signal in each subblock is carried out so that the peak-to-peak variations in the block fit the dynamic range of the quantizer.

3) *Optimized non-uniform quantization* [9], [10]: Rather than adopting uniform scalar quantization, the quantization levels can be optimized as a function of the statistics of the baseband signal by means of standard strategies such as the Lloyd-Max algorithm.

4) *Noise shaping* [11]: Due to the correlation of successive baseband samples, predictive, or noise shaping, quantization techniques based on a feedback filter can be beneficial to reduce the rate of optimized quantization.

5) *Lossless compression* [12][2]: Any residual correlation among successive quantized baseband samples, possibly after predictive quantization, can be further leveraged by entropy coding techniques that aim at reducing the rate down to the entropy of the digitized signal.

As a rule of thumb, compressed CPRI techniques are seen to reduce the fronthaul rate by a factors around 3 [6].

B.2 Alternative Functional Splits

In order to obtain further fronthaul rate reductions by means of point-to-point compression techniques, alternative functional splits to the conventional C-RAN implementation need to be explored [6], [7]. To this end, some baseband functionalities at the physical (PHY) layer , or Layer 1, can be implemented at the RU, rather than at the CU, such as frame synchronization, FFT/IFFT or resource demapping. Note that, while we focus here on the PHY layer, we discuss different functional splits at Layer 2 in Section V.

A first solution at the PHY layer prescribes the implementation of frame synchronization and FFT in the uplink and of the IFFT in the downlink at the RU (see demarcation "A" in Fig. 2). The rest of the baseband functionalities, such as channel decoding/encoding, are instead performed at the CU. This functional split enables the signal to be quantized in the frequency domain, that is, after the FFT in the uplink and prior to the IFFT in the downlink. Given that the signal has a lower peak-to-average ratio (PAPR) in the frequency domain, particularly in the LTE

[2]Reference [12] in fact considers time-domain modulation and not OFDM but the principle is the same discussed here.

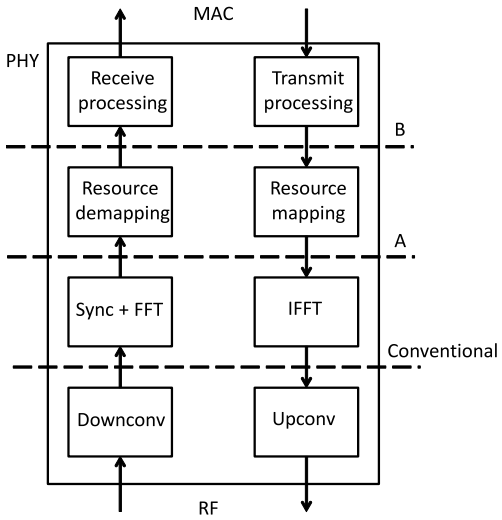


Fig. 2. Alternative functional splits of the physical layer between CU and RU.

downlink, the number of bits per sample can be reduced at a minor cost in terms of signal-to-quantization-noise ratio. The experiments in [6] do not demonstrate, however, very significant rate gains with this approach.

A more promising approach implements also resource demapping for the uplink and resource mapping for the downlink at the RU (see demarcation "B" in Fig. 2). For the uplink, this implies that the RU can deconstruct the frame structure and distinguish among the different physical channels multiplexed in the resource blocks. As a result, the RU can apply different quantization strategies to distinct physical channels, e.g., by quantizing more finely channels carrying higher-order modulations. More importantly, in the case of lightly loaded frames, unused resource blocks can be neglected. This approach was shown in [6], [13] to lead to compression ratios of the order of up to 30, hence an order of magnitude larger than with compressed CPRI, in the regime of small system loads. A similar approach is also implemented in the field trials reported in [14].

C. Network-Aware Compression

The solutions explored so far to address the problem of the excessive fronthaul capacity required by the C-RAN architecture have been based on point-to-point quantization and compression algorithms. Here we revisit the problem by taking a more fundamental viewpoint grounded in network information theory. As it will be discussed below, this network-aware perspective on the design of fronthaul transmission strategies has the potential to move significantly beyond the limitations of point-to-point approaches towards the network information-theoretic optimal performance.

C.1 Uplink

We start by analyzing the uplink. When taking a network-level perspective, a key observation is that the signals (1) received by different RUs are correlated due to the fact that they represent noisy versions of the same signals $\mathbf{x}^{\mathrm{ul}}$. This correlation is expected to be particularly significant for dense networks – an important use case for the C-RAN architecture. Importantly, the fact that the received signals are correlated can be leveraged by the RUs by implementing *distributed source coding* algorithms, which have optimality properties in network information theory (see, e.g., [15] for an introduction).

The key idea of distributed source coding can be easily explained with reference to the problem of compression or quantization with side information at the receiver's side. Specifically, given that the signals received by different RUs are correlated, once the CU has recovered the signal of one RU, that signal can be used as *side information* for the decompression of the signal of another RU. This side information enables the second RU to reduce the required fronthaul rate with no penalty on the accuracy of the quantized signal. This process can be further iterated in a decision-feedback-type loop, whereby signals that have been already decompressed can be used as side information to alleviate the fronthaul requirements for the RUs whose signals have yet to be decompressed.

The coding strategy to be implemented at the RUs to leverage the side information at the receiver is known in information theory as *Wyner-Ziv coding*. Note that Wyner-Ziv coding does not require the RU to be aware of the side information available at the CU but only of the correlation between the received signal and the side information.

Distributed source coding, or Wyner-Ziv coding, was demonstrated in a number of theoretical papers, including [16]–[19], to offer significant potential performance gains. For example, in [20], it was shown via numerical results to nearly double the edge-cell throughput for fixed average spectral efficiency and fronthaul capacities when implemented in a single macrocell overlaid with multiple smaller cells.

The implementation of Wyner-Ziv coding, including both quantization and compression, can leverage the mature state of the art on modern source coding (see, e.g., [21] and references therein). Nevertheless, an important issue that needs to be tackled is the need to inform each RU about the correlation between the received signal and the side information. This correlation depends on the channel state information of the involved RUs and can be provided by the CU to the RU. More practically, the CU could simply inform the RU about which particular quantizer/compressor to apply among the available algorithms in a codebook of possible choices. The design of such codebook and of rules for the selection of specific quantizers/compressors is an open research problem. A discussion on Wyner-Ziv coding using information theoretic arguments can be found in Section IV.

C.2 Downlink

In the downlink, the traditional solution consisting of separate fronthaul quantizers/compressors is suboptimal, from a network information-theoretic viewpoint, based on a different principle, namely that of multivariate compression (see, e.g., [15] for an introduction).

To introduce this principle, we first observe that the quantization noise added by fronthaul quantization can be regarded as a source of interference that affects the UEs' reception. With conventional point-to-point solutions, the CU has little control on

this interfering signal given that the quantization/compression mapping is done separately for each RU. Multivariate, or joint, compression of the signals of all RUs overcomes this problem by enabling the shaping of the quantization regions for the vector of transmitted signals of all RUs. As a result, multivariate quantization/compression makes it possible to control the distribution of the quantization noise across multiple RUs in a similar way as precoding allows to shape the transmission of the useful signals across all the connected RUs.

The idea of multivariate compression for the C-RAN downlink was proposed in [22]. Moreover, recognizing that the quantization noise can be seen as an additional form of interference, reference [22] proposes to perform a joint design of precoding and multivariate compression using an information theoretic formulation. It was shown in [20] that multivariate compression yields performance gains that are comparable to distributed source coding for the uplink.

At a practical level, the implementation of multivariate compression hinges on the availability of channel state information at the CU, which is to be expected, and requires the CU to inform the RU about the quantization levels corresponding to each RU. As for the uplink, the resulting design issues are interesting open problems. Information theoretic considerations on multivariate compression can be found in Section IV.

IV. BASEBAND PROCESSING

As discussed in Section I, one of the key advantages of the C-RAN architecture is that it provides a platform for joint baseband signal processing across the multiple RUs in both uplink and downlink. Such a cooperative network is often referred to as a network MIMO or CoMP [23]. Joint transmission and reception across the RUs allow the possibility for pre-compensation and subtraction of interference across the cells. As inter-cell interference is the dominant performance limiting factor in cellular networks, the C-RAN architecture can achieve significantly higher data rates than conventional cellular networks.

As also seen, a key consideration in the design of cooperative coding strategies for C-RAN is the capacity limit of the fronthaul. In this section, we elaborate on the mathematical modeling of the compression process for a C-RAN system with limited fronthaul and illustrate the effect of compression on baseband signal processing by adopting an information theoretic framework.

The feasibility of cooperative joint signal processing in the C-RAN architecture depends crucially on the ability of the RUs to obtain instantaneous channel state information (CSI) and to precisely synchronize with each other. In the uplink, timing differences can, in theory, be corrected for in the digital domain, but downlink synchronization is imperative so that signals transmitted by the different RUs are received synchronously at the intended UE so as to achieve the cooperative beamforming effect. The rest of the section assumes the availability of CSI and the ability for the RUs to synchronize, and focuses on baseband beamforming design for inter-cell interference mitigation. Furthermore, as described in Section II, a flat-fading channel model is assumed in order to illustrate the fundamental coding strategies in both uplink and downlink.

A. Uplink

As introduced in Section II, the C-RAN architecture consists of RUs that are partitioned into clusters, where each cluster consists of N_R RUs and is responsible for jointly decoding the transmitted information from N_U UEs. Let M_R and M_U be the number of antennas in each of the RUs and the UEs respectively. The discrete-time baseband uplink C-RAN channel model can be written as (1), which can be further detailed as

$$\begin{bmatrix} \mathbf{y}_1^{\text{ul}} \\ \mathbf{y}_2^{\text{ul}} \\ \vdots \\ \mathbf{y}_{N_R}^{\text{ul}} \end{bmatrix} = \begin{bmatrix} \mathbf{H}_{1,1}^{\text{ul}} & \mathbf{H}_{1,2}^{\text{ul}} & \cdots & \mathbf{H}_{1,N_U}^{\text{ul}} \\ \mathbf{H}_{2,1}^{\text{ul}} & \mathbf{H}_{2,2}^{\text{ul}} & \cdots & \mathbf{H}_{2,N_U}^{\text{ul}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{H}_{N_R,1}^{\text{ul}} & \mathbf{H}_{N_R,2}^{\text{ul}} & \cdots & \mathbf{H}_{N_R,N_U}^{\text{ul}} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1^{\text{ul}} \\ \mathbf{x}_2^{\text{ul}} \\ \vdots \\ \mathbf{x}_{N_U}^{\text{ul}} \end{bmatrix} + \begin{bmatrix} \mathbf{z}_1^{\text{ul}} \\ \mathbf{z}_2^{\text{ul}} \\ \vdots \\ \mathbf{z}_{N_R}^{\text{ul}} \end{bmatrix} \tag{3}$$

where the noise terms $\mathbf{z}_i^{\text{ul}}$ are assumed to be additive white complex Gaussian vectors with variance σ_{ul}^2 on each of its components. The goal of joint uplink processing is to utilize the received signals from all the RUs, i.e., $\{\mathbf{y}_1^{\text{ul}}, \mathbf{y}_2^{\text{ul}}, \cdots, \mathbf{y}_{N_R}^{\text{ul}}\}$, to jointly decode $\mathbf{x}_1^{\text{ul}}, \mathbf{x}_2^{\text{ul}}, \cdots, \mathbf{x}_{N_U}^{\text{ul}}$.

The discussion in this section is restricted to a single-hop fronthaul topology, where each RU j is connected to the CU via a digital link of finite capacity. If the fronthaul link capacity had been unlimited, the uplink channel model would have been akin to a multiple-access channel with all the multiple antennas across all the RUs being regarded as to form a single receiver. In this case, well-known strategies such as linear receive beamforming and successive interference cancellation (SIC) could be directly applied across the RUs to approach the best achievable rates of such a multiple-access channel. In designing the receive beamformers across the RUs, the minimum mean-square error (MMSE) beamforming strategy, or a simpler zero-forcing beamforming strategy, can be used, while treating multiuser interference as part of the background noise.

The coding strategy is considerably more complicated when the finite-capacity constraints of the fronthaul links are taken into consideration. Toward this end, as seen in Section III, the RUs must compress its observations and send a compressed version of its IQ samples to the CU. From an information-theoretic viewpoint, the effect of compression can be modeled as additional quantization noises (see, e.g., [24]). For example, if a simple scalar uniform quantization scheme with L quantization levels is used for each I and Q component on each receive antenna, the quantization noise is approximately a uniform random variable within the range $[-L/2, L/2]$. Assuming that the maximum amplitude of the received signal in each of the antennas is within the interval $[-M/2, M/2]$, the amount of fronthaul capacity needed to support such uniform quantization is then $2\log_2(M/L)$ bits per sample, where the factor 2 accounts for the I and Q components. Note that the setting of the value L provides a tradeoff between the fronthaul capacity and the achievable rates. Intuitively, a coarser quantization, i.e., larger L, results in larger quantization noise, thus lower achievable rates, but also less fronthaul. Conversely, finer quantization, i.e.,

smaller L, results in higher achievable rates, but also requires more fronthaul capacity.

To capture such a tradeoff mathematically, and also to account for the fact that vector quantization both across the antennas for each RU and across multiple samples can be used, instead of scalar quantization, for higher quantization efficiency, it is convenient to make the additional assumption that the quantization noise can be modeled as an independent Gaussian process, i.e.,

$$\hat{\mathbf{y}}_j^{\mathrm{ul}} = \mathbf{y}_j^{\mathrm{ul}} + \mathbf{q}_j^{\mathrm{ul}} \tag{4}$$

where $\mathbf{q}_j \sim \mathcal{CN}(\mathbf{0}, \mathbf{Q}_j^{\mathrm{ul}})$ and $\mathbf{Q}_j^{\mathrm{ul}}$ is an $M_R \times M_R$ covariance matrix representing the compression of the received signals across M_R antennas at the jth RU. With this model of the compression process, the overall achievable rate can now be readily written down as a function of the fronthaul capacity.

To this end, assume that each of the UEs $\mathbf{x}_i^{\mathrm{ul}}$ transmits using a Gaussian codebook $\mathcal{CN}(\mathbf{0}, \mathbf{\Sigma}_i^{\mathrm{ul}})$ with possibly multiple data streams per user. In case of linear MMSE receive beamforming across the RUs, the achievable rate for the ith UE can be expressed as:

$$\begin{aligned} R_i^{\mathrm{linear,ul}} &\le I(\mathbf{x}_i^{\mathrm{ul}}; \hat{\mathbf{y}}_1^{\mathrm{ul}} \cdots \hat{\mathbf{y}}_{N_R}^{\mathrm{ul}}) \qquad (5) \\ &= \log \frac{\left| \sum_{j=1}^{N_R} \mathbf{H}_{\mathcal{N}_R,j}^{\mathrm{ul}} \mathbf{\Sigma}_j^{\mathrm{ul}} (\mathbf{H}_{\mathcal{N}_R,j}^{\mathrm{ul}})^H + \mathbf{Q}_{\mathcal{N}_R}^{\mathrm{ul}} + \sigma_{\mathrm{ul}}^2 \mathbf{I} \right|}{\left| \sum_{j \neq i} \mathbf{H}_{\mathcal{N}_R,j}^{\mathrm{ul}} \mathbf{\Sigma}_j^{\mathrm{ul}} (\mathbf{H}_{\mathcal{N}_R,j}^{\mathrm{ul}})^H + \mathbf{Q}_{\mathcal{N}_R}^{\mathrm{ul}} + \sigma_{\mathrm{ul}}^2 \mathbf{I} \right|}. \qquad (6) \end{aligned}$$

When successive interference cancellation is used, assuming without loss of generality a decoding order of the UEs as $1, 2, \cdots, N_U$, the achievable rate for the ith user can instead be expressed as:

$$\begin{aligned} R_i^{\mathrm{SIC,ul}} &\le I(\mathbf{x}_i^{\mathrm{ul}}; \hat{\mathbf{y}}_1^{\mathrm{ul}} \cdots \hat{\mathbf{y}}_{N_R}^{\mathrm{ul}} | \mathbf{x}_1^{\mathrm{ul}}, \cdots, \mathbf{x}_{i-1}^{\mathrm{ul}}) \qquad (7) \\ &= \log \frac{\left| \sum_{j=i}^{N_R} \mathbf{H}_{\mathcal{N}_R,j}^{\mathrm{ul}} \mathbf{\Sigma}_j^{\mathrm{ul}} (\mathbf{H}_{\mathcal{N}_R,j}^{\mathrm{ul}})^H + \mathbf{Q}_{\mathcal{N}_R}^{\mathrm{ul}} + \sigma_{\mathrm{ul}}^2 \mathbf{I} \right|}{\left| \sum_{j=i+1}^{N_R} \mathbf{H}_{\mathcal{N}_R,j}^{\mathrm{ul}} \mathbf{\Sigma}_j^{\mathrm{ul}} (\mathbf{H}_{\mathcal{N}_R,j}^{\mathrm{ul}})^H + \mathbf{Q}_{\mathcal{N}_R}^{\mathrm{ul}} + \sigma_{\mathrm{ul}}^2 \mathbf{I} \right|}. \qquad (8) \end{aligned}$$

In both cases, $\mathbf{H}_{\mathcal{N}_R,j}^{\mathrm{ul}}$ denotes the jth block-column of the matrix $\mathbf{H}^{\mathrm{ul}}$, i.e., the collective channel from UE j to all the RUs. Note that the quantization process simply results in an additional noise term in the rate expression, with the noise covariance matrix defined as $\mathbf{Q}_{\mathcal{N}_R}^{\mathrm{ul}} = \mathrm{diag}(\mathbf{Q}_1^{\mathrm{ul}}, \cdots, \mathbf{Q}_{N_R}^{\mathrm{ul}})$.

The above expression implicitly assumes that the quantization process is done using point-to-point techniques, i.e., independently at each RU (see Section III). In this case, the amount of fronthaul capacity needed to support such quantization at RU j can be expressed based on rate-distortion theory as (see, e.g., [15])

$$\begin{aligned} C_j^{\mathrm{indep,ul}} &\ge I(\mathbf{y}_j^{\mathrm{ul}}; \hat{\mathbf{y}}_j^{\mathrm{ul}}) \qquad (9) \\ &= \log \frac{\left| \sum_{i=1}^{N_U} \mathbf{H}_{ji}^{\mathrm{ul}} \mathbf{\Sigma}_i^{\mathrm{ul}} (\mathbf{H}_{ji}^{\mathrm{ul}})^H + \sigma_{\mathrm{ul}}^2 \mathbf{I} + \mathbf{Q}_j^{\mathrm{ul}} \right|}{\left| \mathbf{Q}_j^{\mathrm{ul}} \right|}. \qquad (10) \end{aligned}$$

Alternatively, as discussed in Section III, Wyner-Ziv compression can be used to take advantage of the fact that the compression of RUs can be done sequentially so that the compressed signals of earlier RUs can act as the decoder side information for the compression of later RUs. Assuming without loss of generality a decompression order of $1, 2, \cdots, N_R$ for the RUs, the fronthaul capacity constraint with Wyner-Ziv compression can be shown to be

$$\begin{aligned} &C_j^{\mathrm{WZ,ul}} \\ &\ge I(\mathbf{y}_j^{\mathrm{ul}}; \hat{\mathbf{y}}_j^{\mathrm{ul}} | \hat{\mathbf{y}}_1^{\mathrm{ul}}, \cdots, \hat{\mathbf{y}}_{j-1}^{\mathrm{ul}}) \qquad (11) \\ &= \log \frac{\left| \mathbf{H}_{\mathcal{J}_j \mathcal{N}_U}^{\mathrm{ul}} \mathbf{\Sigma}_{\mathcal{N}_U}^{\mathrm{ul}} (\mathbf{H}_{\mathcal{J}_j \mathcal{N}_U}^{\mathrm{ul}})^H + \sigma_{\mathrm{ul}}^2 \mathbf{I}_{\mathcal{J}_j} + \mathbf{Q}_{\mathcal{J}_j}^{\mathrm{ul}} \right|}{\left| \mathbf{H}_{\mathcal{J}_{j-1} \mathcal{N}_U}^{\mathrm{ul}} \mathbf{\Sigma}_{\mathcal{N}_U}^{\mathrm{ul}} (\mathbf{H}_{\mathcal{J}_{j-1} \mathcal{N}_U}^{\mathrm{ul}})^H + \sigma_{\mathrm{ul}}^2 \mathbf{I}_{\mathcal{J}_{j-1}} + \mathbf{Q}_{\mathcal{J}_{j-1}}^{\mathrm{ul}} \right| \cdot \left| \mathbf{Q}_j^{\mathrm{ul}} \right|} \qquad (12) \end{aligned}$$

where we use the notations $\mathcal{J}_j = \{1, 2, \cdots, j\}$ and $\mathcal{N}_U = \{1, 2, \cdots, N_U\}$, and use $\mathbf{H}_{\mathcal{J}_j \mathcal{N}_U}^{\mathrm{ul}}$ to denote the block-submatrix of $\mathbf{H}^{\mathrm{ul}}$ with indices taken from $\mathcal{J}_j$ and $\mathcal{N}_U$. Likewise, $\mathbf{\Sigma}_{\mathcal{N}_U}^{\mathrm{ul}}$ denotes a block-diagonal matrix with block-diagonal entries $\mathbf{\Sigma}_1^{\mathrm{ul}}, \cdots, \mathbf{\Sigma}_{N_U}^{\mathrm{ul}}$; and a similar definition applies to $\mathbf{Q}_{\mathcal{J}_{j-1}}^{\mathrm{ul}}$.

In summary, the uplink rate expressions (6) for linear receive beamforming and (8) for successive interference cancellation provide information theoretical characterizations of the uplink C-RAN capacity limit subject to fronthaul capacity constraints with either independent per-link quantization (10) or Wyner-Ziv quantization (12). These expressions implicitly assumes the use of capacity and rate-distortion achieving codes, but the performance with practical codes can also be easily obtained by incorporating gap factors in the expressions (see e.g., [25]). Also implicit in the expressions is the decoding strategy at the CU of decoding the compression codewords at the RUs first and then the transmitted codewords from the UEs. Such a strategy has information theoretical justification [16], but we remark that this is not the only possible decoding strategy (see e.g., [26], [27]). Furthermore, as mentioned, the implementation of this strategy assumes MMSE beamforming across the RUs. The beamforming coefficients typically need to be designed centrally at the CU as functions of the global CSI.

The achievable rate characterization points to the possibility that the transmit covariance of the UEs and the quantization noise covariance at the RUs may be jointly designed in order to maximize the overall system performance. For example, a weighted rate-sum maximization problem may be formulated over user scheduling, power control, transmit beamforming at the UEs, the quantization noise covariance matrices at the RUs, and possibly the successive compression and successive interference cancellation orders at the CU. Various forms of this problem have appeared in the literature [17]–[19], [28]. The implementation of the solutions to such an optimization, however, depends on the feasibility of adaptive coding, modulation, and quantization codebooks, according to CSI, scheduling, and user rates. As a first step for implementing C-RAN, fixed-rate scalar uniform quantization is more likely to be used with quantization level set according to the dynamic range of the analog-to-digital convertors and the subsequent fronthaul capacity limits (see Section III). In fact, as shown in [19], uniform quantization noise level is approximately optimal under suitable high signal-to-noise ratio (SNR) conditions. In this case, the quantization noise simply becomes additional background noise to be taken

into consideration when designing scheduling, power control, and receive beamforming strategies.

B. Downlink

In the downlink C-RAN architecture, baseband processing at the CU involves linear beamforming or non-linear techniques such as dirty paper coding that aim at ensuring that the signals transmitted by the RUs are received at the UE in such a way that interference is minimized. If the fronthaul links between the RUs and the CU have infinite capacities, the downlink C-RAN becomes a broadcast channel and standard network information theoretic results apply [15]. The situation is instead more involved when the fronthaul links have finite capacities. In this case, as seen, after the CU forms the beamformed signals to be transmitted by the RUs, as functions of the user data and CSI, such signals need to be compressed before they can be sent to the RUs.

Mathematically, the discrete-time baseband downlink C-RAN channel model can be written as (2), or more specifically as

$$\begin{bmatrix} \mathbf{y}_1^{\mathrm{dl}} \\ \mathbf{y}_2^{\mathrm{dl}} \\ \vdots \\ \mathbf{y}_{N_U}^{\mathrm{dl}} \end{bmatrix} = \begin{bmatrix} \mathbf{H}_{1,1}^{\mathrm{dl}} & \mathbf{H}_{1,2}^{\mathrm{dl}} & \cdots & \mathbf{H}_{1,N_R}^{\mathrm{dl}} \\ \mathbf{H}_{1,2}^{\mathrm{dl}} & \mathbf{H}_{2,2}^{\mathrm{dl}} & \cdots & \mathbf{H}_{2,N_R}^{\mathrm{dl}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{H}_{N_U,1}^{\mathrm{dl}} & \mathbf{H}_{N_U,2}^{\mathrm{dl}} & \cdots & \mathbf{H}_{N_U,N_R}^{\mathrm{dl}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{x}}_1^{\mathrm{dl}} \\ \hat{\mathbf{x}}_2^{\mathrm{dl}} \\ \vdots \\ \hat{\mathbf{x}}_{N_R}^{\mathrm{dl}} \end{bmatrix} + \begin{bmatrix} \mathbf{z}_1^{\mathrm{dl}} \\ \mathbf{z}_2^{\mathrm{dl}} \\ \vdots \\ \mathbf{z}_{N_U}^{\mathrm{dl}} \end{bmatrix} \tag{13}$$

where $\mathbf{z}_i^{\mathrm{dl}}$ is the additive white complex Gaussian vector with zero mean and variance σ_{dl}^2 on each of its components. Note that in a time-division duplex (TDD) system, the reciprocity of the uplink and downlink channels would mean that $\mathbf{H}_{i,j}^{\mathrm{dl}} = (\mathbf{H}_{j,i}^{\mathrm{ul}})^T$.

The transmit signals $\hat{\mathbf{x}}_j^{\mathrm{dl}}$ are quantized versions of the beamformed signals $\mathbf{x}_j^{\mathrm{dl}}$. The quantization process can be modeled as the addition of quantization noises as discussed above, yielding

$$\hat{\mathbf{x}}_j^{\mathrm{dl}} = \mathbf{x}_j^{\mathrm{dl}} + \mathbf{q}_j^{\mathrm{dl}}. \tag{14}$$

An interesting aspect of downlink quantization is that, in contrast to uplink, where the quantization encoding in each RU is necessarily independent, in the downlink the encoding operation is done centrally at the CU, and thus *correlated* quantization noises can be introduced. Such a compression scheme is called multivariate compression, first introduced in the C-RAN context in [22], as discussed in Section III.

Let $\mathbf{s}_i^{\mathrm{dl}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma}_i^{\mathrm{dl}})$ be the beamformed signal intended for the ith UE to be transmitted across the RUs, which may contain multiple data streams. The eigenvectors of $\mathbf{\Sigma}_i^{\mathrm{dl}}$ are the transmit beamformers over the RUs. As the desired transmit signal across the RUs is a combination of the intended signals for all the N_U UEs, i.e., $\mathbf{x}^{\mathrm{dl}} = \sum_{i=1}^{N_U} \mathbf{s}_i^{\mathrm{dl}}$, the transmit signal across the RUs is therefore

$$\begin{bmatrix} \mathbf{x}_1^{\mathrm{dl}} \\ \mathbf{x}_2^{\mathrm{dl}} \\ \vdots \\ \mathbf{x}_{N_R}^{\mathrm{dl}} \end{bmatrix} \sim \mathcal{CN}\left(\mathbf{0}, \sum_{i=1}^{N_U} \mathbf{\Sigma}_i^{\mathrm{dl}}\right). \tag{15}$$

In order to describe the quantization process, we rewrite components of the transmit covariance matrix corresponding to each of the RUs separately as

$$\sum_{i=1}^{N_U} \mathbf{\Sigma}_i^{\mathrm{dl}} = \begin{bmatrix} \mathbf{S}_{1,1}^{\mathrm{dl}} & \mathbf{S}_{1,2}^{\mathrm{dl}} & \cdots & \mathbf{S}_{1,N_R}^{\mathrm{dl}} \\ \mathbf{S}_{1,2}^{\mathrm{dl}} & \mathbf{S}_{2,2}^{\mathrm{dl}} & \cdots & \mathbf{S}_{2,N_R}^{\mathrm{dl}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_{N_R,1}^{\mathrm{dl}} & \mathbf{S}_{N_R,2}^{\mathrm{dl}} & \cdots & \mathbf{S}_{N_R,N_R}^{\mathrm{dl}} \end{bmatrix} \tag{16}$$

and also the quantization noise covariance as

$$\begin{bmatrix} \mathbf{q}_1^{\mathrm{dl}} \\ \mathbf{q}_2^{\mathrm{dl}} \\ \vdots \\ \mathbf{q}_{N_R}^{\mathrm{dl}} \end{bmatrix} \sim \mathcal{CN}\left(\mathbf{0}, \begin{bmatrix} \mathbf{Q}_{1,1}^{\mathrm{dl}} & \mathbf{Q}_{1,2}^{\mathrm{dl}} & \cdots & \mathbf{Q}_{1,N_R}^{\mathrm{dl}} \\ \mathbf{Q}_{1,2}^{\mathrm{dl}} & \mathbf{Q}_{2,2}^{\mathrm{dl}} & \cdots & \mathbf{Q}_{2,N_R}^{\mathrm{dl}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{Q}_{N_R,1}^{\mathrm{dl}} & \mathbf{Q}_{N_R,2}^{\mathrm{dl}} & \cdots & \mathbf{Q}_{N_R,N_R}^{\mathrm{dl}} \end{bmatrix}\right) \tag{17}$$

where $\mathbf{S}_{i,j}^{\mathrm{dl}}$ and $\mathbf{Q}_{i,j}^{\mathrm{dl}}$ are $N_R \times N_R$ matrices.

If we use point-to-point fronthaul compression, the quantization noises are independent and hence uncorrelated, i.e., $\mathbf{Q}_{i,j}^{\mathrm{dl}} = \mathbf{0}$ for $i \neq j$, and the fronthaul capacity needed to generate $\hat{\mathbf{x}}_j^{\mathrm{dl}}$ is simply:

$$C_j^{\mathrm{indep,dl}} \geq I(\mathbf{x}_j^{\mathrm{dl}}; \hat{\mathbf{x}}_j^{\mathrm{dl}}) \tag{18}$$

$$= \log \frac{\left|\mathbf{S}_{jj}^{\mathrm{dl}} + \mathbf{Q}_{jj}^{\mathrm{dl}}\right|}{\left|\mathbf{Q}_{jj}^{\mathrm{dl}}\right|}. \tag{19}$$

If we instead utilize multivariate compression, as described in Section III, to generate correlated quantization noises, extra fronthaul capacity would be needed. In a dual manner as in the uplink Wyner-Ziv coding case, we restrict attention here to the performance achievable using successive encoding [22]. Without loss of generality, let the encoding order of RUs be $1, 2, \cdots, N_R$. By simplifying the information theoretical expressions of [22], it can be shown that the required fronthaul capacity can be expressed as follows:

$$C_j^{\mathrm{multi,dl}} \geq I(\mathbf{x}_j^{\mathrm{dl}}; \hat{\mathbf{x}}_j^{\mathrm{dl}}) + I(\mathbf{q}_j^{\mathrm{dl}}; \mathbf{q}_1^{\mathrm{dl}}, \cdots, \mathbf{q}_{j-1}^{\mathrm{dl}}) \tag{20}$$

$$= \log \frac{\left|\mathbf{S}_{jj}^{\mathrm{dl}} + \mathbf{Q}_{jj}^{\mathrm{dl}}\right|}{\left|\mathbf{Q}_{jj}^{\mathrm{dl}}\right|} + \log \frac{\left|\mathbf{Q}_{jj}^{\mathrm{dl}}\right|}{\left|\mathbf{Q}_{jj}^{\mathrm{dl}} - \mathbf{Q}_{j\mathcal{J}_{j-1}}^{\mathrm{dl}} (\mathbf{Q}_{\mathcal{J}_{j-1}\mathcal{J}_{j-1}}^{\mathrm{dl}})^{-1} \mathbf{Q}_{\mathcal{J}_{j-1}j}^{\mathrm{dl}}\right|}, \tag{21}$$

where $\mathcal{J}_{j-1} = \{1, \cdots, j-1\}$ and $\mathbf{Q}_{\mathcal{J}_{j-1}\mathcal{J}_{j-1}}^{\mathrm{dl}}$ denotes the submatrix of the quantization covariance indexed by the subscripts, and likewise for $\mathbf{Q}_{\mathcal{J}_{j-1}j}^{\mathrm{dl}}$. Although generating correlated quantization noises requires extra fronthaul capacity, as elaborated on in Section III, multivariate compression brings the advantage that the effective total noise at the UEs may be lowered as the correlated quantization noises at the RUs can potentially cancel each other after going through the channel, thus improving the overall user rates for the system.

When multiuser interference is treated as noise, the achievable downlink rate can be expressed as a function of the quanti-

zation noise covariance as

$$R_i^{\mathrm{linear,dl}} \tag{22}$$
$$\leq I(\mathbf{s}_i^{\mathrm{dl}};\mathbf{y}_i^{\mathrm{dl}})$$
$$= \log \frac{\left|\sum_{j=1}^{N_U} \mathbf{H}_{i,\mathcal{N}_R}^{\mathrm{dl}} \left(\mathbf{\Sigma}_j^{\mathrm{dl}} + \mathbf{Q}_{\mathcal{N}_R}^{\mathrm{dl}}\right) (\mathbf{H}_{i,\mathcal{N}_R}^{\mathrm{dl}})^H + \sigma_{\mathrm{dl}}^2 \mathbf{I}\right|}{\left|\sum_{j\neq i} \mathbf{H}_{i,\mathcal{N}_R}^{\mathrm{dl}} \left(\mathbf{\Sigma}_j^{\mathrm{dl}} + \mathbf{Q}_{\mathcal{N}_R}^{\mathrm{dl}}\right) (\mathbf{H}_{i,\mathcal{N}_R}^{\mathrm{dl}})^H + \sigma_{\mathrm{dl}}^2 \mathbf{I}\right|}, \tag{23}$$

where $\mathcal{N}_R = \{1,\cdots,N_R\}$, $\mathbf{H}_{j,\mathcal{N}_R}^{\mathrm{dl}}$ is the jth block-row of the channel matrix $\mathbf{H}^{\mathrm{dl}}$, i.e., the collective channel from the RUs to UE j, and $\mathbf{Q}_{\mathcal{N}_R}^{\mathrm{dl}}$ is the quantization noise covariance matrix across the N_R RUs. When dirty-paper coding is used in the downlink, multiuser interference can be pre-subtracted. Assuming without loss of generality a successive precoding order of UE $1,2,\cdots,N_U$, the achievable rate for the ith user can be expressed as:

$$R_i^{\mathrm{DPC,dl}} \tag{24}$$
$$\leq I(\mathbf{s}_i^{\mathrm{dl}};\mathbf{y}_i^{\mathrm{dl}}|\mathbf{s}_1^{\mathrm{dl}},\cdots,\mathbf{s}_{i-1}^{\mathrm{dl}})$$
$$= \log \frac{\left|\sum_{j=i}^{N_U} \mathbf{H}_{i,\mathcal{N}_R}^{\mathrm{dl}} \left(\mathbf{\Sigma}_j^{\mathrm{dl}} + \mathbf{Q}_{\mathcal{N}_R}^{\mathrm{dl}}\right) (\mathbf{H}_{i,\mathcal{N}_R}^{\mathrm{dl}})^H + \sigma_{\mathrm{dl}}^2 \mathbf{I}\right|}{\left|\sum_{j=i+1}^{N_U} \mathbf{H}_{i,\mathcal{N}_R}^{\mathrm{dl}} \left(\mathbf{\Sigma}_j^{\mathrm{dl}} + \mathbf{Q}_{\mathcal{N}_R}^{\mathrm{dl}}\right) (\mathbf{H}_{i,\mathcal{N}_R}^{\mathrm{dl}})^H + \sigma_{\mathrm{dl}}^2 \mathbf{I}\right|}. \tag{25}$$

In summary, the rate expressions (23) for linear transmit beamforming and (25) for dirty-paper coding provide information theoretical characterization of the downlink C-RAN capacity limit subject to fronthaul capacity constraints with either per-link quantization (19) or multivariate quantization (21). As above, the use of capacity and rate-distortion achieving codes is assumed, but the expressions can be easily modified to account for practical coding and compression methods. These rate characterizations again provide the possibility that the transmit covariance intended for each UE and the quantization noise covariance at the RUs may be jointly designed in order to maximize the overall system performance. For example, a weighted sum-rate maximization problem may be formulated over user scheduling, downlink power control, transmit beamformers, and the quantization covariance setting at the RUs. Although this joint system-level design problem is non-convex and fairly difficult to solve, algorithms capable of achieving local optimum solutions have been devised for some forms of this problem in [22], [25].

As in the uplink, the implementation of such solutions would require the use of adaptive modulation, adaptive quantization, and the availability of global CSI. Thus again, a first step for implementation of downlink C-RAN is likely to involve simpler beamforming designs (such as zero-forcing) and scalar fixed quantizers designed according to the per-antenna power constraints and the fronthaul capacity limits.

C. Alternative Functional Splits

The discussion above assumes the standard C-RAN implementation in which the RUs are remote antenna heads tasked with compression only and not with encoding and decoding of the UE data. As seen in Section III, this functional split is preferred in C-RAN in order to make the RUs as simple as possible, but it is not the only possible strategy. In terms of baseband processing, in the downlink, the CU may opt to share user data directly with the RUs, instead of sharing the compressed version of the beamformed signals. The resulting replication of the UE data at multiple RUs yields an inefficient use of the fronthaul link capacity when the cooperation cluster size is large enough. However, data-sharing can be effective when the fronthaul capacity is limited, i.e., when the cluster size is relatively small (see, e.g., [29]). The optimization of the cooperation cluster is an interesting problem, which has been dealt with extensively in the literature [30], [31], [33]–[37].

V. MEDIUM ACCESS CONTROL

In the previous sections, we have discussed cooperative techniques at the PHY layer that leverage the C-RAN architecture. As seen, these methods require the deployment of new infrastructure, including RUs and fronthaul link with tight capacity and latency constraints. It is, however, also of interest for mobile network operators to find solutions that reuse the existing infrastructure with the goal of cost-efficiently enhancing it with some centralized RAN functionalities. The implementation of an RU-CU functional split at Layer 2 is a promising candidate solution to achieve this goal, as it has been reported to drastically reduce the fronthaul requirements – up to factor 20 depending on the system configuration [38], [39] – while still allowing for centralization gains by means of coordinated radio resource management (RRM). This section briefly reviews challenges and opportunities related to Layer 2 functional splitting.

Fig. 3 shows several functional split options for Layer 2 of the radio protocol stack. Note that, in the following, we use terminology based on the 3GPP LTE specifications. Since other technologies such as IEEE 802.16 (WiMAX) have a similar radio architecture with functional equivalents, the discussion here applies in principle for them as well. The Layer 2 is structured in sub-layers as follows [40]:

- Medium access control (MAC): this sub-layer is responsible for multiplexing and scheduling of control and user plane data into logical channels and transport blocks, and for hybrid automatic repeat request (HARQ) aimed at fast recovery from block errors;
- Radio link control (RLC): this higher sub-layer is tasked with the segmentation of user data for the MAC scheduler, with buffering and with the ARQ protocol for improved link reliability;
- Packet data convergence protocol (PDCP): this sub-layer, placed on top of the RLC sub-layer, is responsible for ciphering and integrity protection, for data forwarding aimed at hand-over, and for header compression on small data packets (e.g., voice data);
- Radio resource control (RRC): this sub-layer implements the control-plane protocol for radio link management and configuration, including measurements, admission control, and hand-over control.

An important aspect of Layer 2 sub-layers is the classification into *synchronous* and *asynchronous* protocols, which refers to the timing of the corresponding frame building process at the base station: synchronous protocols need to deliver data (e.g., transport blocks in case of MAC) within the LTE transmis-

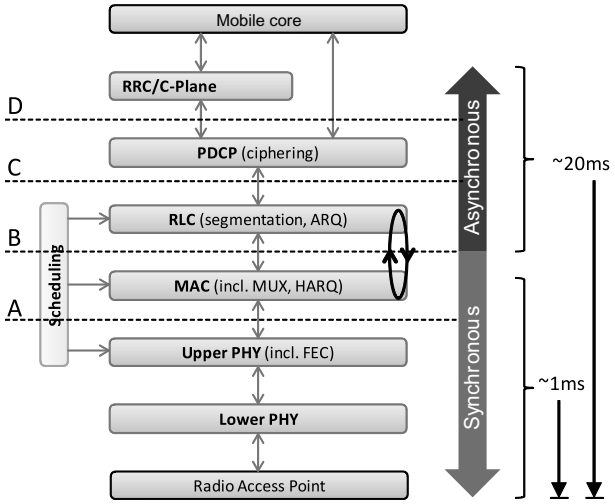


Fig. 3. Functional split options on RAN at Layer 2.

sion time intervals (TTI) of 1 ms and hence have more stringent requirements on latency and jitter than the *asynchronous* sub-layers, whose latency requirements are of the order of 20 ms.

The type and corresponding latency requirement, along with advantages and disadvantages, of different functional splits at Layer 2 are summarized in Table 1. In the rest of this section, we provide a more detailed discussion on these aspects.

A. Constraints and Requirements

The RU-CU split of Layer 2 functionalities is subject to specific constraints and requirements. To start, we observe that the fronthaul capacity requirements are less critical than for PHY-layer functional splits. In particular, the additional overhead of the control plane, e.g., signalling radio bearers, RRC messages, and protocol headers at Layer 2 adds approximately 10 % to the overall bandwidth requirement as driven by user plane traffic, yielding for LTE to up to a theoretical maximum overall fronthaul rate of, e.g., 150 Mbps in downlink for a 20 MHz FDD system with 2 transmit antennas [38]. Note that this rate corresponds to worst-case traffic conditions, which should be considered when dimensioning the system, whereby the available radio resources are fully occupied and the highest modulation and coding scheme (MCS) (ordinal number 28 in the 3GPP specifications) is used.

A first set of constraints arises from the fact that some functions are located in a single protocol layer, such as segmentation in RLC and ciphering in PDCP, while others span several protocol layers. This imposes implementation constraints on potential functional splits that involve the latter type of protocols, because information exchange and consequently signalling between RU and CU would be required if the corresponding functional split were implemented. An example is scheduling, which encompasses the upper PHY (for assigning transport blocks to resource blocks), the MAC sub-layer (for multiplexing and QoS scheduling) and the RLC sub-layer (for extracting the required number of bytes from corresponding buffers). In particular, the MAC scheduler needs to know the buffer occupancy at the RLC sub-layer in order to extract the selected number of bytes from the RLC radio bearer buffers according to the available radio resources and scheduled calculation. This process needs to be completed in a fraction of the TTI of 1 ms and involves a bi-directional information exchange, as indicated in Fig. 3. As a consequence, barring a re-consideration of the protocol stack design of LTE, split B in Fig. 3 can be in practice ruled out as a potential candidate due to the discussed tight integration of MAC and RLC sub-layers.

Other implementation constraints are determined by feedback loops involving the mobile device and by the related use of timers and procedures based on time-out events of some protocols. This is, for instance, the case for HARQ and ARQ, respectively at the MAC and RLC sub-layers, as well as for hand-over and connection control functions at the RRC sub-layer. Specifically, the HARQ feedback loop is the main constraining timer for all functional split options at Layer 2. As illustrated in Fig. 4, the mobile device side expects an acknowledgement (negative or positive) in sub-frame $n + 4$ counting from the sub-frame of the transmission in uplink [41]. This imposes a limitation of below 3 ms on the round-trip time budget. As also indicated in the figure, this budget includes the round-trip transmission over the fronthaul as well as the processing and frame building at the CU. Assuming that sufficient processing power is available at the CU, this leads to a maximum tolerable fronthaul one-way latency of approximately 1 ms – buffering for jitter not included. This requirement constitutes a challenge for functional split A. In the literature, only few papers have addressed this challenge so far, e.g., [6], [42]–[44]. Above functional split A, timing requirements are less stringent. Specifically, ARQ and RRC timers are configurable, but in order to ensure the performance of key indicators such as hand-over failure and residual block error rate, a maximum latency in the range of 10 to 20 ms for split options C and D should be assumed.

B. Centralization Gains

Having discussed the drawbacks related to requirements and implementation constraints of different functional splits at Layer 2, we now elaborate on their relative advantages in terms of centralization gains. Centralization gains can be classified into *multiplexing gains*, which depend on the statistical properties of aggregated traffic and processing demand at the CU, and *coordination gains*, which are due to coordinated radio resource management and control for a cluster of RUs. Note that the latter implies that the system implementation is capable of exchanging required information between protocol entities in the CU, which requires interfaces between functional entities (e.g., APIs).

Multiplexing gains for traffic aggregation apply mainly to the interfaces from the CU to the mobile core network. The reason is that the fronthaul needs to be capable of carrying the maximum possible throughput per RU for most splits. For splits C and D on asynchronous protocols, the fronthaul could be dimensioned also on a higher quartile of the bandwidth distribution (e.g., 99 % quartile) with the risk of introducing some additional delay, or even additional packet losses [45]. Coordination gains instead depend on the split option and the corresponding centralized functions, and two main cases can be distinguished:

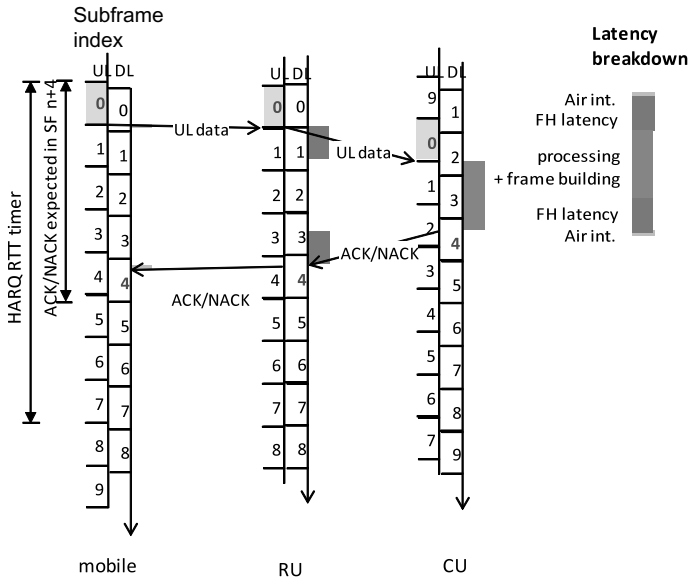


Fig. 4. HARQ timing in LTE systems.

- *Coordinated RRM*: this form of centralization includes scheduling, inter-cell interference coordination (ICIC), cell-based discontinuous transmission (DTX) and other techniques where the CU decides on the allocation and transmission of resources on a per-frame basis. The corresponding gains can be attained for split option A and below. For functional splits higher in the protocol stack, a centralized, more long term RRM approach is possible but would require additional signalling between the CU and the MAC entities in the RU (see, e.g., [46]–[48]).
- *Centralized RRC*: this type of centralization amounts to the coordination of admission control, load balancing, hand-over parametrization and related self-organizing network (SON) functions which are executed on longer time ranges. These gains are possible for all split options, including split C and D.

A comprehensive overview of potential coordination gains depending on the functional split option is provided in [45]. Finally, there are some implicit gains such as centralization of ciphering in split option C, which implies that data is already protected for transport to the RUs and therefore does not need additional, and costly, transport layer security.

C. Summary

In summary, although various split options on Layer 2 are possible in current cellular systems, the design of the protocol stack points strongly towards two main candidates, namely splitting below MAC (split A), which benefits from centralized RRM at the price of high requirements on backhaul latency, and a split between PDCP and RLC (split C), which is cost efficient and is already standardized as dual connectivity in LTE [40]. As mentioned, Table 1 provides an overview of the pros and cons of the four considered split options.

Table 1. Summary of Layer 2 functional split options.

Split	Type	Fronthaul latency req.	Centralization gains	Pros	Cons
D	asynch.	20 ms	Admission control, load balancing, SON functionality	Low req. on fronthaul and computational resources	C-plane centralization only
C	asynch.	20 ms	D + moderate processing gains	Low req., fronthaul ciphering included	Coordinated scheduling would require additional signaling
B	synch.	$\ll 1$ ms	None	None	Very high req. on latency, additional signalling required
A	synch.	< 1 ms	C + centralized ICIC, scheduling	High RRM gains possible	Higher req. on latency

VI. RADIO RESOURCE MANAGEMENT

As discussed in the previous section, in a C-RAN with a functional split below the level A in Fig. 3, RRM may be carried out at the CU in a centralized fashion for the cluster of connected RUs. This centralized optimization is based on the available information at the CU, including queue state information, CSI and topological information about the fronthaul network. Due to the limitations of the fronthaul network and the need for possibly large-scale centralized optimization, RRM optimization in C-RANs offers significant technical challenges that are briefly reviewed, along with the state of the art on existing solutions, in this section. Specifically, we first discuss in subsection VI-A the static RRM problem that aims at maximizing performance metrics such as weighted sum rate in a given frame. Then, in subsection VI-B, we elaborate on the more general RRM problem of allocating resources across successive frames in a dynamic fashion by adapting to the available CSI and queue state information. As it will be discussed, solutions to the static problem often serve as components of the techniques addressing the dynamic scenario (see, e.g., [49]).

A. Static RRM

The static RRM problem amounts to the maximization of performance criteria such weighted sum-rate on a per-frame basis based on the available CSI. Below, we first discuss fully centralized solutions and then party decentralized approaches that leverage game-theoretic tools.

A.1 Centralized Optimization

The optimization of typical performance criteria, such as weighted sum-rate, amounts to non-convex, and possibly combinatorial, optimization problems with respect to the resource variables of interest, including downlink beamforming, uplink user association and RU clustering. Furthermore, these prob-

lems typically involves constraints that account for the limited fronthaul resources, such as the requirement to activate only a subset of RUs. We briefly review some approaches and solutions in the following.

Non-convexity with respect to the downlink beamforming variables is caused by the presence of inter-cell, or inter-RU, interference. This can be generally dealt with in various ways, most notably via successive convex approximation methods (see, e.g., [50]) and via techniques based on Fenchel-duality arguments or, equivalently, on the weighted minimum mean square error (WMMSE) method [51]. Instead, the mentioned fronthaul constraints are often formulated by introducing an l_0-norm regularization term in the objective function that enforces a penalty which is proportional to the number of active RUs, or, in other words, to the sparseness of the RU activation vector. To transform the corresponding non-convex problems into convex ones, standard l_1-norm approximation methods can be used to ensure sparsity of the resulting solution, or, more generally, mixed l_1/l_p-norm approximation techniques can be adopted to induce group sparsity (see, e.g., [35]).

The approaches discussed above have been applied in the context of C-RAN in [31], [52] with the aim of minimizing energy consumption; in [53] for joint power and antenna selection optimization; in [35] for weighted sum-rate maximization; and in [54] for joint downlink precoding and uplink user-RU association optimization.

A.2 Decentralized Optimization

The centralized optimization discussed above requires the availability of CSI at the CU, which may impose a significant burden on the fronthaul, especially for large-scale C-RANs. To reduce this overhead, one may resort to decentralized solutions whereby the RUs self-organize into clusters based only on collected local information. To this end, the framework of coalition games can be adopted to develop cluster formation algorithm. This was proposed in reference [55], which uses as utility function of a cluster the total data rate, and leverages a merge-split algorithm to obtain a stable cluster partition. A related work is [56], in which interference from a legacy base station is considered that is coordinated with a coexisting C-RAN by means of a contract-based approach. Here, the proposed scheme aims at maximizing the utility of the C-RAN while preserving the performance of the legacy base station.

B. Dynamic RRM

While the solutions reviewed above operate on a per-frame basis, in practice, RRM needs to operate across multiple frames and to be adaptive to the time-varying conditions of the channel on the RAN and to the state of the queues, as illustrated in Fig. 5. Dynamic RRM solutions that tackle this problem are reviewed in the following, by focusing on the two prominent approaches based on Markov decision processes (MDP) and Lyapunov optimization.

B.1 Markov Decision Processes

To elaborate, the system state of a C-RAN in a given frame can be generally characterized by the current CSI and queue state information, which we denote as $\chi(t) = [\mathbf{H}(t), \mathbf{Q}(t)]$,

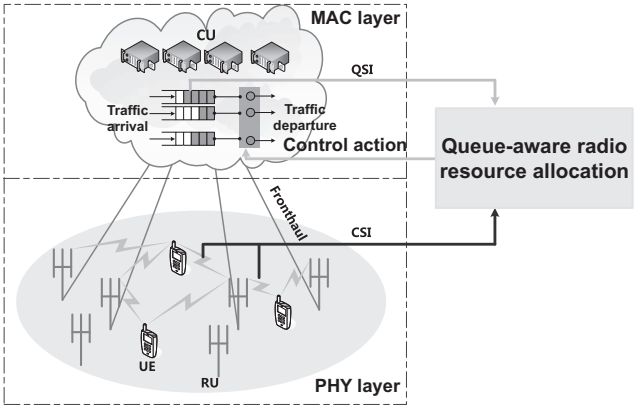


Fig. 5. An illustration of dynamic RRM in C-RANs.

where t represents the frame index, $\mathbf{H}(t)$ is the current CSI and the vector $\mathbf{Q}(t)$ describes the state of the queues. Under a Markovian model for the state $\chi(t)$, the dynamic RRM problem can be modeled as a finite or infinite horizon average cost MDP. Under proper technical conditions, this problem can be in principle solved by tackling the Bellman equation. However, this approach incurs the curse of dimensionality, since the number of system states grows exponentially with the number of traffic queues maintained by the centralized CU. To overcome this problem, the methods of approximate MDP, stochastic learning, and continuous-time MDP could be used (see, e.g., [57]). The problem is even more pronounced in the absence of full state information, in which case the framework of Partially Observable MDPs (POMDPs), with its added complexity, needs to be considered.

A dynamic RRM solution that operates at both PHY and MAC layers has been proposed in [58] in the presence of imperfect CSI at the CU for the downlink by leveraging the POMDP framework. This reference proposes to reduce the complexity of the resulting solution by describing the trajectory of traffic queues by means of differential equations and hence in terms of a continuous-time MDP. In so doing, the value functions can be easily calculated using calculus, hence substantially reducing the computational burden.

B.2 Lyapunov Optimization

Lyapunov optimization provides another systematic approach for dynamic RRM optimization in C-RANs. Lyapunov optimization-based techniques are able to stabilize the queues hosted at the CUs while additionally optimizing some time-averaged performance metric [49]. The approach hinges on the minimization of the one-step conditional Lyapunov drift-plus-penalty function

$$E[L(\mathbf{Q}(t+1)) - L(\mathbf{Q}(t))|\mathbf{Q}(t)] + VE[g(t)|\mathbf{Q}(t)] \quad (26)$$

where $L(\mathbf{Q}(t)) = \frac{1}{2}\sum_{i\in\mathcal{I}} Q_i(t)^2$ is the Lyapunov function obtained by summing the squares of the queues' occupancies, $g(t)$ is the system cost at slot t and V is a adjustable control parameter.

The dynamic RRM problem of network power consumption minimization by means of joint RU activation and downlink

beamforming was studied in [60], [61] by leveraging the Lyapunov optimization framework. As shown therein, the resulting algorithm requires the solution of a static penalized weighted sum rate problem at each frame, which may be tackled as discussed above. Reference [62] includes also congestion control in the problem formulation and derives corresponding solutions based on Lyapunov optimization.

VII. SYSTEM-LEVEL CONSIDERATIONS

In this section, we provide a brief discussion on network architectures implementing C-RAN systems. We first discuss the basic architecture in subsection VII-A and then briefly cover more advanced solutions in subsection VII-B.

A. C-RAN Network Architecture

Fig. 6 shows the basic architecture of a C-RAN system, which consists of the access, fronthaul, backhaul and packet core segments. In this architecture, the cell sites in the access network are connected to the cloud center, or CU, through fronthaul links. As discussed in Section III and Section V, the RUs may implement different functionalities at Layer 1 and Layer 2. In the most basic deployment, the RUs only perform RF operations, such as frequency up/down conversion, sampling and power amplification. In this case, the RUs contain the antennas and RF front-end hardware as well as the fronthaul interface software, e.g., CPRI, to communicate with the CU. Possible additional functionalities at Layer 1 and Layer 2 necessitate extra hardware and software modules at the RUs to coordinate and communicate with the CU (see, e.g., [39]).

The transport technology used in the fronthaul affects, and depends on, parameters such as cost, latency and distance between the radio sites and the Cloud Center. Fiber-optic and microwave links are the leading transport media for fronthauling, encompassing a large percentage of existing C-RAN developments [39]. Dedicated fiber solutions between the Cloud Center and RUs provide significant performance in terms of data rate and latency, but have been met with limited deployment due to cost associated with it. Optical transport networks (OTN), along with wavelength division multiplexing networks (WDM), provide high spectral efficiency by enabling fiber sharing among different cell sites with bidirectional transmission between the RUs and Cloud Center.

In the CU, multiple baseband units may be collocated that coordinate for the execution of the operations virtualized by the RUs in the access networks. A key design challenge for Cloud Centers is the development of cost-effective and high performance baseband pooling platforms, which may use, as further discussed in [2], either digital signal processing (DSP) or general purpose processors (GPP) technologies. In addition to processing related to network access, the CU is also responsible for the interaction with the backbone network, e.g., Evolved Packet Core in LTE, via the backhaul (see Fig. 6). Current C-RAN technology solutions and trials mainly consider the fronthaul and backhaul segments separately, as in the recent platforms presented by Huawei and Ericsson [39]. In this case, fronthaul signals, which adhere to the serial CPRI transport format, are translated into Ethernet-based backhaul transport signals at the Cloud Center for transmission on the backhaul.

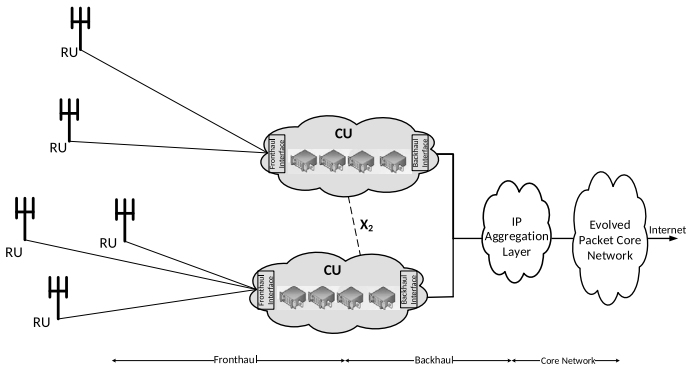


Fig. 6. C-RAN system architecture.

Due to baseband pooling and to the constraints on the I/Q data transmission in the fronthaul segment, the C-RAN architecture does not utilize the X2 interface to the extent of existing LTE systems. However, as shown in Fig. 6, C-RAN clustering methods across multiple CUs are also considered to leverage coordinated transmission for wider areas [2]. In this regard, we observe that, even though inter-CU coordination cannot be as efficient as intra-RU coordination due to latency and capacity limitations in the fronthaul, multiple network management techniques such as time/frequency resource silencing techniques may be implemented across CUs.

B. Next-Generation C-RAN Network Architecture

In the state-of-the-art C-RAN architecture discussed above, it is generally quite complex, costly and inefficient to manage flexibly and dynamically the resources of the fronthaul, backhaul and core network segments. This is due to the heterogeneous technologies used for the corresponding network devices and their control elements. Recent advances in software defined networking (SDN) technology, with its successful implementations such as OpenFlow, motivate the utilization of SDN network management tools for C-RAN deployments in order to overcome this limitation. Fig. 7 demonstrates a reference architecture that targets SDN-based unified network operation and transport mechanisms across the fronthaul, backhaul, and core network segments of a C-RAN architecture [63], [64].

In this architecture, a unified SDN-based control plane interfaces with the C-RAN network elements through dedicated control channels. Virtualized functions at the RUs, described in Section III (cf. Fig. 3) and identified in Fig. 7 for short as fA,...,fD, are dynamically coordinated by the SDN controller which assigns them to the corresponding nodes in the network. The function assignment procedure is based on network and link level abstractions at the network elements, which are populated through southbound and northbound interfaces and conveyed to the SDN controller. The abstracted parameters conveyed from the network elements to the controller through the northbound interface may include a wide range of inputs including link level conditions at the fronthaul such as bandwidth, signal to interference noise ratio, delay, etc. Similarly, the SDN controller, utilizing the southbound interface, conveys the configuration and execution information of the underlying virtual functions that

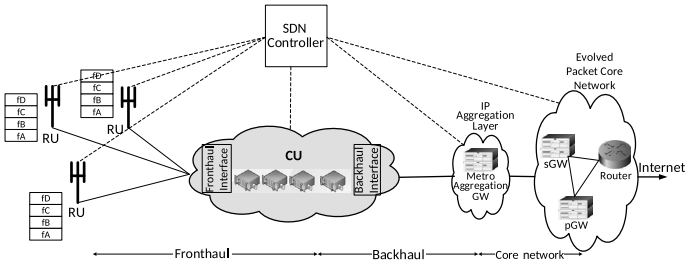


Fig. 7. Flexible C-RAN architecture with fronthaul and backhaul segments.

are dynamically allocated per network element, e.g., RU, at a given network instance [63].

VIII. STANDARDIZATION

The discussed potential performance gains and reduction in operating and maintenance cost offered by the C-RAN technology have resulted in significant industrial research and development efforts over the last decade. Similar to other incumbent telecommunication technologies and their life-cycles, the initial phase of C-RAN development was mostly led by the individual contributions and demonstrations of leading companies such as China Mobile [66], Huawei [67], Ericsson [39], Nokia Siemens Networks [68], and others. The subsequent, and ongoing, standardization efforts on C-RAN aim at developing a compatible fronthaul technology and its interfaces at the RUs and CU that enable multi-vendor operations. We briefly review below some of these activities.

As discussed in Section III, CPRI is currently the most widely deployed industry alliance standard that defines the specifications for the interface between the radio equipment controller (REC) and the radio equipment (RE). CPRI defines a digitized I/Q transmission interface that supports serial, bidirectional and constant rate transmission on the fronthaul. The standard includes specifications for control plane, including strict synchronization and low-latency transmission via configured CPRI packetization, and for data plane [5].

Open Base Station Architecture Initiative [69] and Open Radio Interface [70] are other competing standards and industry associations that define interfaces and functional descriptions for the base station transceiver.

Aiming at providing multi-tenancy support and dynamic functional allocation, hardware and functional virtualization have been discussed under the umbrella of NFV ISG in ETSI [71]. Specifically, NFV ISG in ETSI targets a framework for telecom network virtualization that is directly applicable to the C-RAN architecture and aims at reducing the cost of deployment, at enabling multi-tenancy operation, and at allowing for easier operating and maintenance procedures.

IX. CONCLUDING REMARKS

This article has provided a short review of the state of the art and of ongoing activities in the industry and academia around the C-RAN technology. We have highlighted practical and theoretical aspects at Layer 1, including fronthaul compression and baseband processing; at Layer 2, with an emphasis on RU-CU functional splits; and at higher layers, including radio resource management. We have also discussed network architecture considerations and standardization efforts. Throughout the article, a tension has been emphasized between the two trends of virtualization, which prescribes wireless access nodes with only RF functionalities and entails significant capacity and latency requirements on the fronthaul architecture; and edge processing, which instead involves the implementation of a subset of Layer 1 and possibly also of Layer 2 functions at the edge nodes so as to reduce delays and alleviate architectural constraints. Ongoing activities point to solutions that find a balance between these two trends by means of flexible RAN and fronthaul/ backhaul technologies which allow the adaptation of the network operation to traffic type and system conditions.

REFERENCES

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

[2] A. Checko *et al.*, "Cloud RAN for mobile networks - A technology overview," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 405–426, 2015.

[3] H. Al-Raweshidy and S. Komaki, *Radio over fiber technologies for mobile communications networks*. Artech House, 2002.

[4] C. Lu, M. Berg, E. Trojer, P.-E. Eriksson, K. Laraqui, O. V. Tidbl, and H. Almeida, "Connecting the dots: small cells shape up for high-performance indoor radio," *Ericsson Review*, 2014.

[5] Ericsson AB, Huawei Technologies, NEC Corporation, Alcatel Lucent, and Nokia Siemens Networks, "Common public radio interface (CPRI); interface specification," CPRI specification v5.0, Sept. 2011.

[6] U. Dotsch, M. Doll, H. P. Mayer, F. Schaich, J. Segel, and P. Sehier, "Quantitative analysis of split base station processing and determination of advantageous architectures for LTE." *Bell Labs Tech. J.*, vol. 18, no. 1, pp. 105–128, 2013.

[7] D. Wubben, P. Rost, J. Bartelt, M. Lalam, V. Savin, M. Gorgoglione, A. Dekorsy, and G. Fettweis, "Benefits and impact of cloud computing on 5G signal processing: Flexible centralization through cloud-RAN." *IEEE Signal Process. Mag.*, no. 31, pp. 35–44, 2014.

[8] Integrated Device Technology, Inc., "Front-haul compression for emerging C-RAN and small cell networks," Apr. 2013.

[9] D. Samardzija, J. Pastalan, M. MacDonald, S. Walker, and R. Valenzuela, "Compressed transport of baseband signals in radio access networks," *IEEE Trans. Wireless Commun.*, vol. 11, no. 9, pp. 3216–3225, Sept. 2012.

[10] B. Guo, W. Cao, A. Tao, and D. Samardzija, "CPRI compression transport for LTE and LTE-A signal in C-RAN," in *Proc. Int. ICST Conf. CHINACOM*, 2012, pp. 843–849.

[11] K. F. Nieman and B. L. Evans, "Time-domain compression of complex-baseband LTE signals for cloud radio access networks," in *Proc. IEEE GlobalSIP*, 2013, pp. 1198–1201.

[12] A. Vosoughi, M. Wu, and J. R. Cavallaro, "Baseband signal compression in wireless base stations," in *Proc. IEEE GLOBECOM*, Dec. 2012, pp. 4505–4511.

[13] J. Lorca and L. Cucala, "Lossless compression technique for the fronthaul of LTE/LTE-advanced cloud-RAN architectures," in *Proc. IEEE WoWMoM*, June 2013, pp. 1–9.

[14] S. Grieger, S. Boob, and G. Fettweis, "Large scale field trial results on frequency domain compression for uplink joint detection," in *Proc. IEEE GLOBECOM*, 2012, pp. 1128–1133.

[15] A. E. Gamal and Y.-H. Kim, *Network information theory*, Cambridge University Press, 2011.

[16] A. Sanderovich, O. Somekh, H. V. Poor, and S. Shamai (Shitz), "Uplink macro diversity of limited backhaul cellular network," *IEEE Trans. Inf. Theory*, vol. 55, no. 8, pp. 3457–3478, Aug. 2009.

[17] A. del Coso and S. Simoens, "Distributed compression for MIMO coordinated networks with a backhaul constraint," *IEEE Trans. Wireless Commum.*, vol. 8, no. 9, pp. 4698–4709, Sept. 2009.

[18] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai (Shitz), "Robust and efficient distributed compression for cloud radio access networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 2, pp. 692–703, Feb. 2013.

[19] L. Zhou and W. Yu, "Optimized backhaul compression for uplink cloud radio access network," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1295–1307, June 2014.

[20] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai (Shitz), "Performance evaluation of multiterminal backhaul compression for cloud radio access networks," in *Proc. CISS*, Princeton, NJ, Mar. 19–21, 2014.

[21] M. J. Wainwright, "Sparse graph codes for side information and binning," *IEEE Signal Process. Mag.*, vol. 24, no. 5, pp. 47–57, Sept. 2007.

[22] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai (Shitz), "Joint precoding and multivariate backhaul compression for the downlink of cloud radio access networks," *IEEE Trans. Signal Process.*, vol. 61, no. 22, pp. 5646–5658, Nov. 2013.

[23] D. Gesbert, S. Hanly, H. Huang, S. Shamai (Shitz), O. Simeone, and Wei Yu, "Multi-Cell MIMO cooperative networks: A new look at interference," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1380–1408, Dec. 2010.

[24] A. Gersho and R. M. Gray, *Vector quantization and signal compression*, Kluwer Acad. Press, 1992.

[25] P. Patil, B. Dai, and W. Yu, "Performance comparison of data-sharing and compression strategies for cloud radio-access networks," in *Proc. EUSIPCO*, 2015.

[26] L. Zhou and W. Yu, "Uplink multicell processing with limited backhaul via per-base-station successive interference cancellation," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 10, pp. 1981–1993, Oct. 2013.

[27] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai (Shitz), "Joint decompression and decoding for cloud radio access networks," *IEEE Signal Process. Lett.*, vol. 20, no. 5, pp. 503–506, May 2013.

[28] Y. Zhou and W. Yu, "Optimized beamforming and backhaul compression for uplink MIMO cloud radio-access networks," in *Proc. IEEE GLOBECOM*, 2014.

[29] J. Kang, O. Simeone, J. Kang, and S. Shamai (Shitz), "Fronthaul compression and precoding design for C-RANs over ergodic fading channels," to appear in *IEEE Trans. Veh. Technol.*

[30] Jian Zhao, T. S. Quek, and Zhongding Lei, "Coordinated multipoint transmission with limited backhaul data transfer," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2762–2775, June 2013.

[31] Yuanming Shi, Jun Zhang, and K. B. Letaief, "Group sparse beamforming for green Cloud-RAN," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2809–2823, May 2014.

[32] A. Liu and V. Lau, "Joint power and antenna selection optimization in large cloud radio access networks," *IEEE Trans. Signal Process.*, vol. 62, no. 5, pp. 1319–1328, Mar. 2014.

[33] Fuxin Zhuang and V. K. N. Lau, "Backhaul limited asymmetric cooperation for MIMO cellular networks via semidefinite relaxation," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 684–693, Feb. 2014.

[34] Jun Zhang, Runhua Chen, J. G. Andrews, A. Ghosh, and R. W. Heath, "Networked MIMO with clustered linear precoding," *IEEE Trans. Wireless Commun.*, vol. 8, no. 4, pp. 1910–1921, Apr. 2009.

[35] B. Dai and W. Yu, "Sparse beamforming and user-centric clustering for downlink Cloud Radio Access Network," *IEEE Access*, vol. 2, pp. 1326–1339, 2014.

[36] R. Zakhour, and D. Gesbert, "Optimized data sharing in multicell MIMO with finite backhaul capacity," *IEEE Trans. Signal Process.*, vol. 59, no. 12, pp. 6102–6111, Dec. 2011.

[37] P. Marsch and G. Fettweis, "Uplink CoMP under a constrained backhaul and imperfect channel knowledge," *IEEE Trans. Wireless Commun.*, vol. 10, no. 6, pp. 1730–1742, June 2011.

[38] Small Cell Forum, "Small cell virtualization functional splits and use cases", White Paper, June 2015.

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

[40] 3GPP, "TS 36.300 V12.5.0; Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description," Mar. 2015.

[41] 3GPP, "TS 36.213 V10.4.0; Physical layer procedures," Dec. 2011.

[42] P. Rost and A. Prasad, "Opportunistic Hybrid ARQ–Enabler of Centralized-RAN over non-ideal backhaul", *IEEE Wireless Commun. Lett.*, vol. 3, no. 5, pp. 481–484, 2014

[43] S. Khalili and O. Simeone, "Uplink HARQ for Distributed and Cloud RAN via Separation of Control and Data Planes," available on arXiv (arXiv:1508.06570).

[44] Q. Han, C. Wang, M. Levorato, and O. Simeone, "On the effect of fronthaul latency on ARQ in C-RAN systems," submitted, available on arXiv (arXiv:1510.07176).

[45] INFSO-ICT-317941 iJOIN, "Deliverable D5.3; Final definition of iJOIN architecture", Apr. 2015.

[46] INFSO-ICT-317941 iJOIN, "Deliverable D3.3; Final definition and evaluation of MAC and RRM approaches for RANaaS and a joint backhaul/access design," Apr. 2015.

[47] R. Fritzsche, P. Rost, G. Fettweis, "Robust proportional fair scheduling with imperfect channel state information", accepted for publication in *IEEE Trans. Wireless Commun.*, 2015.

[48] E. Pateromichelakis, M. Shariat, A. Quddus, and R. Tafazolli, "Graph-based multicell scheduling in OFDMA-based small cell networks", *IEEE Access*, vol. 2, pp.897–908, 2014.

[49] M. Neely, *Stochastic network optimization with application to communication and queueing systems*, Morgan & Claypool Publishers, 2010.

[50] G. Scutari, F. Facchinei, L. Lampariello, and P. Song, "Distributed methods for constrained nonconvex multi-agent optimization-Part I: Theory," arXiv:1410.4754.

[51] Q. Shi, M. Razaviyayn, Z. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sept. 2011.

[52] J. Tang, W. Tay, and T. Quek, "Cross-layer resource allocation with elastic service scaling in cloud radio access network," *IEEE Trans. Wireless Commun.*, vol. PP, no. 99, p. 1, May. 2015.

[53] A. Liu and V. Lau, "Joint power and antenna selection optimization in large cloud radio access networks," *IEEE Trans. Signal Process.*, vol. 62, no. 5, pp. 1319–1328, Mar. 2014.

[54] S. Luo, R. Zhang, and T.J. Lim, "Downlink and uplink energy minimization through user association and beamforming in C-RAN," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 494–508, Jan. 2015.

[55] Z. Zhao *et al.*, "Cluster formation in cloud radio access networks: performance analysis and algorithms design," in *Proc. ICC*, London, UK, June 2015, pp. 1–6.

[56] M. Peng *et al.*, "Contract-based interference coordination in heterogeneous cloud radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1140–1153, Mar. 2015.

[57] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming,* Wiley-Interscience, 2005.

[58] J. Li, M. Peng, A. Cheng, Y. Yu, and C. Wang, "Resource allocation optimization for delay-sensitive traffic in fronthaul constrained cloud radio access networks," *IEEE Syst. J.*, pp. 1–12, Nov. 2014.

[59] J. Li, M. Peng, A. Cheng, and Y. Yu, "Delay-aware cooperative multipoint transmission with backhaul limitation in cloud-RAN," in *Proc. IEEE ICC*, Sydney, Australia, June. 2014, pp. 665–670.

[60] P. Teseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *J. Opt. Theory App.*, vol. 109, no. 3, pp. 475–494, June 2001.

[61] J. Li, J. Wu, M. Peng, W. Wang, and V. K. N. Lau, "Queue-aware joint remote radio head activation and beamforming for green cloud radio access networks," in *Proc. IEEE GLOBECOM*, San Diego, USA, Dec. 2015.

[62] J. Li, M. Peng, Y. Yu, and A. Cheng, "Dynamic resource optimization with congestion control in heterogeneous cloud radio access networks," *in Proc. IEEE GLOBECOM*, Austin, USA, Dec. 2014, pp. 906–911.

[63] L. Jingchu, T. Zhao, S. Zhou, Y. Cheng, and Z. Niu, "Concert: a cloud-based architecture for next-generation cellular systems," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 14–22, 2014.

[64] A. de la Oliva *et al.*, "Xhaul: Towards an Integrated Fronthaul/Backhaul Architecture in 5G Networks," [Online]. Available: http://eprints.networks.imdea.org/1059/1/Xhaul_Towards_Integrated_Fronthaul_Backhaul_2015_EN.pdf

[65] J. Segel and M. Weldon, "Lightradio portfolio-technical overview," Technology White Paper 1, Alcatel-Lucent.

[66] China Mobile, "C-RAN: the road towards green RAN," White Paper, ver. 2.5, China Mobile Research Institute, Oct. 2011.

[67] Huawei, "Cloud RAN Introduction. The 4th CJK International Workshop Technology Evolution and Spectrum," Sept. 2011.

[68] H. Guan, T. Kolding, and P. Merz, "Discovery of Cloud-RAN," Nokia Siemens Networks, Tech. Rep., April 2010.

[69] "BTS System Reference Document Version 2.0," Open Base Station Architecture Initiative. Nov. 14, 2006. Retrieved August 16, 2013.

[70] ETSI ORI ISG, [Online]. Available: http://portal.etsi.org/tb.aspx?tbid=738&SubTB=738

[71] ETSI NFV ISG, "Network Functions Virtualisation," Dec., 2012, [Online]. Available: http://portal.etsi.org/portal/server.pt/community/NFV/367

[72] "The benefits of Cloud-RAN architecture in mobile network expansion", Fujitsu, White paper, 2015.

[73] Ghebretensae *et al.*, "Transmission solutions and architectures for heterogeneous networks built as C-RANs." in *Proc. International Conference on Communications and Networking in China*, 2012.

[74] O. Simeone, N. Levy, A. Sanderovich, O. Somekh, B. M. Zaidel, H. V. Poor, and S. Shamai (Shitz), "Cooperative wireless cellular systems: an information-theoretic view," *Foundations and Trends in Communications and Information Theory*, vol. 8, nos. 1-2, pp. 1–177, 2012.

[75] S. Hur, T. Kim, D. J. Love, J. V. Krogmeier, T. A. Thomas, and A. Ghosh, "Millimeter wave beamforming for wireless backhaul and access in small cell networks," *IEEE Trans. Commun.*, vol. 61, no. 10, pp. 4391–4403, Oct. 2013.

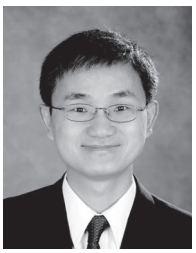
[76] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai (Shitz), "Multi-hop backhaul compression for the uplink of cloud radio access networks," arXiv:1312.7135.
[77] P. Rost *et al.*, "Cloud technologies for flexible 5G radio access networks," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 68–76, Apr. 2014.
[78] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai (Shitz), "Fronthaul compression for cloud radio access networks: Signal processing advances inspired by network information theory," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 69–79, 2014.

Osvaldo Simeone received the M.Sc. degree (with honors) and the Ph.D. degree in Information Engineering from Politecnico di Milano, Milan, Italy, in 2001 and 2005, respectively. He is currently with the Center for Wireless Communications and Signal Processing Research (CWCSPR), New Jersey Institute of Technology (NJIT), Newark, where he is an Associate Professor. His research interests concern wireless communications, information theory, optimization and machine learning. Dr. Simeone is a co-recipient of the 2015 IEEE Communication Society Best Tutorial Paper Award and of the Best Paper Awards at IEEE SPAWC 2007 and IEEE WRECOM 2007. He currently serves as an Editor for IEEE Transactions on Information Theory. He is a Fellow of the IEEE.

Andreas Maeder is a Senior Radio Researcher in Nokia Bell Labs, where he is leading the research on 5G RAN Architecture work. Andreas received his Ph.D. in 2008 from the University of Wuerzburg, Germany. From 2008 to 2015, Dr. Maeder was affiliated with NEC Laboratories Europe, where he was working on next generation mobile networks, including air interface design for IEEE 802.16m and LTE-Advanced, and system architecture, virtualization, and cloudification of mobile network functional components. He was a key contributor and work package lead of the EU FP7 project iJOIN on Cloud RAN and joint access/backhaul optimization. Dr. Maeder was contributing to the standardization of IEEE 802.16m on WirelessMAN-Advanced and IEEE 802.16p on M2M communications, to 3GPP RAN2, and served as rapporteur for the 3GPP System Architecture working group. He is author of numerous standard contributions, conference papers, journal articles, book chapters and more then 40 patents and patent applications.

Mugen Peng received the B.E. degree in Electronics Engineering from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2000, and the Ph.D. degree in Communication and Information Systems from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2005. After the Ph.D. graduation, he joined BUPT, where he has been a Full Professor with the School of Information and Communication Engineering since 2012. In 2014, he was an Academic Visiting Fellow with Princeton University, Princeton, NJ, USA. He leads a Research Group focusing on wireless transmission and networking technologies with the Key Laboratory of Universal Wireless Communications (Ministry of Education), BUPT. His main research areas include in wireless communication theory, radio signal processing, and convex optimizations, with a particular interests in cooperative communication, radio network coding, self-organization networking, heterogeneous networking, and cloud communication. He has authored or co-authored over 40 refereed IEEE journal papers and over 200 conference proceeding papers. Dr. Peng is currently on the Editorial/Associate Editorial Board of the IEEE Communications Magazine, the IEEE ACCESS, IET Communications, the International Journal of Antennas and Propagation, China Communication, and the International Journal of Communications System. He has been the leading Guest Editor of the special issues on the IEEE Wireless Communications Magazine, the International Journal of Antennas and Propagation, and the International Journal of Distributed Sensor Networks. He was a recipient of the 2014 IEEE ComSoc AP Outstanding Young Researcher Award, and the best paper award in GameNets 2014, CIT 2014, ICCTA 2011, ICBNMT 2010, and IET CCWMC 2009. He received the First Grade Award of the Technological Invention Award in the Ministry of Education of China, and the First and Second Grade Award of Scientific and Technical Progress from the China Institute of Communications.

Onur Sahin received the B.Sc. degree in Electrical and Electronics Engineering from Middle East Technical University, Ankara, Turkey, in 2003 and the M.Sc. and Ph.D. degrees in Electrical Engineering from the Polytechnic Institute of New York University, Brooklyn, in 2005 and 2009, respectively. He is currently Staff Research Engineer at the Innovation Labs, InterDigital Inc. and conducts research and development on next generation telecommunication and wireless systems (including 5G and beyond) with particular emphasis on PHY/MAC layer technologies, network and internet architectures with upper layer protocol design, and network information theory. Dr. Sahin has also held technical lead positions at multiple projects on the development of next generation cellular and Wi-FI systems including LTE-A and IEEE 802.11 standards and is currently involved with several European Union funded H2020 projects for 5G and beyond systems. He is the co-author of over 30 peer-reviewed scientific articles (450+ citations), co-inventor of 15 patents and patent applications, and serves as a Guest Editor for Journal of Communication Networks. Dr. Sahin was a visiting scholar at Imperial College London, UK during the Fall 2013 and Spring 2014 semesters. He is the recipient of 2012 and 2015 InterDigital Innovation Awards.



Wei Yu received the B.A.Sc. degree in Computer Engineering and Mathematics from the University of Waterloo, Waterloo, Ontario, Canada in 1997 and M.S. and Ph.D. degrees in Electrical Engineering from Stanford University, Stanford, CA, in 1998 and 2002, respectively. Since 2002, he has been with the Electrical and Computer Engineering Department at the University of Toronto, Toronto, Ontario, Canada, where he is now Professor and holds a Canada Research Chair (Tier 1) in Information Theory and Wireless Communications. His main research interests include information theory, optimization, wireless communications and broadband access networks. Prof. Wei Yu currently serves on the IEEE Information Theory Society Board of Governors (2015-17). He is an IEEE Communications Society Distinguished Lecturer (2015-16). He served as an Associate Editor for IEEE Transactions on Information Theory (2010-2013), as an Editor for IEEE Transactions on Communications (2009-2011), as an Editor for IEEE Transactions on Wireless Communications (2004-2007), and as a Guest Editor for a number of special issues for the IEEE Journal on Selected Areas in Communications and the EURASIP Journal on Applied Signal Processing. He was a Technical Program co-chair of the IEEE Communication Theory Workshop in 2014, and a Technical Program Committee co-chair of the Communication Theory Symposium at the IEEE International Conference on Communications (ICC) in 2012. He was a member of the Signal Processing for Communications and Networking Technical Committee of the IEEE Signal Processing Society (2008-2013). Prof. Wei Yu received a Steacie Memorial Fellowship in 2015, an IEEE Communications Society Best Tutorial Paper Award in 2015, an IEEE ICC Best Paper Award in 2013, an IEEE Signal Processing Society Best Paper Award in 2008, the McCharles Prize for Early Career Research Distinction in 2008, the Early Career Teaching Award from the Faculty of Applied Science and Engineering, University of Toronto in 2007, and an Early Researcher Award from Ontario in 2006. He was named a Highly Cited Researcher by Thomson Reuters in 2014. Prof. Wei Yu is a Fellow of IEEE. He is a registered Professional Engineer in Ontario.