

16

Implementation of Hardware and Software Platforms

Chia-Yu Chang^{1,}, Dario Sabella², David García-Roger³, Dieter Ferling⁴, Fredrik Tillman⁵, Gian Michele Dell'Aera⁶, Leonardo Gomes Baltar², Michael Färber², Miquel Payaró⁷, Navid Nikaein¹, Pablo Serrano⁸, Raymond Knopp¹, Sandra Roger³, Sylvie Maryrargue⁹ and Tapio Rautio¹⁰*

¹ EURECOM, France

² Intel, Germany

³ Universitat Politècnica de València, Spain

⁴ Nokia, Germany

⁵ Ericsson, Sweden

⁶ Telecom Italia, Italy

⁷ CTTC/CERCA, Spain

⁸ Universidad Carlos III de Madrid, Spain

⁹ CEA-LETI, France

¹⁰ VTT, Finland

* Note that the authors are listed in alphabetical order in this chapter

16.1 Introduction

The exponential growth in mobile data traffic, anticipated to reach a 1000-fold increase over the next decade, and the large diversity of applications - ranging from low bit-rate and low power machine-to-machine (M2M) applications to highly interactive and high-resolution entertainment applications - impose a number of distinct technical requirements on hardware (HW) and software (SW) platforms, where all mobile communication and related functionalities are implemented and executed. Among the most prominent requirements, the following must be highlighted: the further improvement of quality of experience (e.g., capacity, latency, resilience) and the energy efficiency, as well as the scalability, modularity, and reconfigurability when multiple radio access technologies (RATs) are considered.

In order to deal with the increasing traffic demand and, especially, with the anticipated heterogeneity of traffic in 5th generation (5G) cellular systems, future networks will require an increase of HW versatility and the ability to operate with increasingly higher bandwidths, especially at millimetre wave (mmWave) bands. Although the software-defined radio (SDR) paradigm and, in general, the

digitally assisted analogue front-end have contributed in this direction, further improvements cannot be made without novel enhancements on the HW itself. Versatile and/or reconfigurable HW components and platforms, which are able to cope with all the functionalities needed and invoked by the SW domain, are thus required to deliver innovative, cost effective and efficient network elements and devices, in order to achieve a successful commercial exploitation and deployment of 5G.

In addition, the control and the management of the heterogeneous HW infrastructure and devices expected in 5G are of utmost importance. In particular, these should guarantee an effective (i.e., rapid, easy and dependable) service development and deployment, as well as the adaptability to very demanding and changing contexts of operation, while maintaining the Quality of Service (QoS) and Quality of Experience (QoE). Thus, programmability and dynamic reconfiguration through interface abstractions or uniform application programming interfaces (API) are required to enable a HW-agnostic operation.

This chapter addresses implementation challenges and a complexity analysis of HW and SW platforms, and is structured as follows: The focus of Section 16.2 is on analogue and mixed signal HW, whereas in Section 16.3 digital HW is addressed. The link between the HW and SW domains is established in Section 16.4, which deals with HW/SW function partitioning aspects. A set of functional requirements for SW platforms that can cope with the challenges outlined in the previous paragraph is described in Section 16.5. In Section 16.6, an example of platform implementation targeting a virtual radio access network (vRAN) or cloud radio access network (CRAN) architecture is provided, before the chapter is summarized in Section 16.7.

16.2 Solutions for Radio Frontend Implementation

16.2.1 Requirements on 5G Radio Frontends

The requirement of significantly increased radio bandwidth for mobile communication asks for exploiting additional spectral resources to the already used frequency bands, as covered in Section 3.4, and, consequently, impacts the components utilized for the air interface.

The operation in several radio bands defined in the 3rd Generation Partnership Project (3GPP) Long Term Evolution (LTE) standard [1] for carrier aggregation between 700 MHz and 3.6 GHz, has been extended to the frequency range of 450 MHz to 6 GHz to include new available radio bands allowing for increased capacity in mobile systems. The concurrent operation in different radio bands increases the HW complexity when duplicating the individual transceiver chains to support each of the bands. Thus, solutions for multiple band transceivers help to decrease the HW complexity by reducing the number of components and have motivated the research on such topics.

An additional increase of operating bandwidth is achieved by exploiting higher frequency bands of centimetre and millimetre wavelength. There, several GHz of spectrum will become available for mobile communications. System architectures and air interfaces are currently being defined and solutions for radio frontends are necessary.

Multi-antenna systems will be used in all frequency ranges to enhance the radio performance, as addressed in detail in Section 11.5. At the lower frequencies, an increased number of multiple-input multiple-output (MIMO) streams is targeted to increase spectral efficiency by exploiting spatial multiplexing. Similar approaches are also considered for mmWave, where multiple antennas are required to enable mobile radio links at these frequencies by overcoming the particular radio propagation

conditions through beamforming and the related antenna gain. These approaches, however, lead to HW complexity that scales (linearly or even non-linearly) with the number of antennas and MIMO streams. Solutions to decrease complexity are hence required to enable their implementation.

An efficient utilization of spectrum is mandatory due to its limited availability. Solutions for more efficient spectrum usage are necessary to increase the mobile data volume for a defined bandwidth. One considered option is to use a certain frequency band simultaneously for transmission and reception, referred to as full-duplex.

A further requirement on implemented HW components is to improve their energy efficiency with the goal of simplifying their integration, due to lower cooling requirements, and decreasing the carbon footprint of wireless communication systems, as detailed in Section 15.3. This requirement implies the utilization of semiconductor technologies with lower power consumption and techniques for energy efficient component operation. It is worth highlighting here again that if the 1000-fold increase in data rate foreseen in 5G would be expected to dissipate the same energy as in 4G systems, this would require improving the energy efficiency per bit by the same factor of 1000, as also discussed in Section 2.3.

Concepts and solutions to meet the requirements for 5G radio frontends are presented in the following sections, which give insights on recent research, while more details are available in [2] and [3].

16.2.2 Multi-band Transceivers

The focus of this section is on bands below 6 GHz. Some aspects related to mmWave bands can be found in Section 16.2.3.2, especially related to multi-antenna operation, which is key at these higher bands.

To decrease the HW complexity, the signals of multiple radio bands are ideally fed through a single transceiver chain. The multiplication of transceiver chains by the number of supported bands and the increase of the number of included components are avoided, reducing size, weight and cost. The approach is based on broadband or multi-band capability of components used for data and frequency conversion, amplification and filtering.

Newly developed radio frequency (RF) data converters facilitate the generation of signals directly at radio frequencies. This includes the digital-to-analogue and frequency conversion. Operating at sampling rates between 9 and 15 Gsamples/s, they show a broadband performance of up to 2 GHz signal bandwidth positioned arbitrarily at up to 6 GHz carrier frequency. The high-speed serial data interface allows the separated transmission of signals for different radio bands enabling an efficient utilization of the interface. This advantage is supported by numerically controlled oscillators (NCO) together with up or down conversion functionalities included in broadband digital-to-analogue converters (DAC) or analogue-to-digital converters (ADC). In this way, data and frequency conversion are provided by a single component for multiple radio bands.

The implementation of the RF signal generation with newest samples of DACs operating at 12 Gsamples/s, (e.g., those available at [4],[5]), targets three-band operation at carrier frequencies around 2.6, 2.8, and 3.5 GHz, as shown in Figure 16-1. Six times 20 MHz signal carriers provide 120 MHz of aggregated operating bandwidth positioned as two adjacent carriers in the three different radio bands. The pre-distorted signals generated in the digital signal processor (DSP) to compensate nonlinear distortions in the power amplifier (PA) show five times larger bandwidth than the wanted signal and are adapted to the bandwidth of the third and fifth order intermodulation distortions caused by amplifiers. Thus, signals of 600 MHz aggregated bandwidth will pass through the component, utilizing

BB: Baseband signal
 DAC: Digital to analog convertor
 DSP: Digital signal processing
 HPA: High power amplifier
 RB: Radio band
 RF: Radio frequency
 SC: Signal carrier
 VGA: Variable gain amplifier

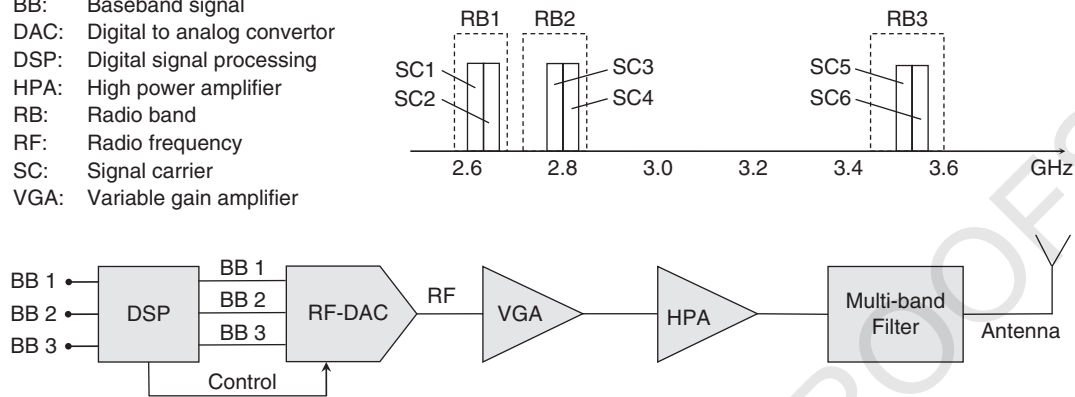


Figure 16-1. Concept of a three-band transmitter with an example of signal carrier positions.

mostly the capacity of the high-speed serial data link, which connects the DSP with the RF-DAC. It supports aggregated multiple complex input data streams up to a maximum complex data rate of 1.5 Gsamples/s. The three individual baseband signals accepted at the digital interface of the RF-DAC are individually configured in gain, converted to RF signals and positioned in the targeted radio bands. These signals feed the amplification stages of the transmitter as shown in the figure.

Broadband amplification is state-of-the-art at low power levels. For higher power, an increase in power efficiency of amplifiers is required to reduce the effort for component cooling and the power consumption of the system. The required energy efficiency stays in contradiction with the wide bandwidth, and a trade-off between output power, bandwidth, and efficiency has to be considered. Accordingly, it is more appropriate to realize multi-band PAs for radio bands positioned in a restricted frequency range to limit the overall amplifier bandwidth, and implement separate transceiver chains for widely spaced bands.

To cover the three envisaged radio bands, a dual-band high power amplifier is designed with a relative bandwidth of 10.9 % (2.6 GHz – 2.9 GHz) and 5.7 % (3.4 – 3.6 GHz) in a Doherty topology [6] with a peak output power of more than 53 dBm. It allows the operation in all three envisaged bands simultaneously. Considering the concurrent operation of six modulated carriers, an average transmission power of up to 38 dBm per carrier is achieved, corresponding to the power level of medium range base stations (BSs). This combination of bandwidth, power level and carrier frequencies is achieved on a Gallium Nitride (GaN) technology by realizing two transistor power elements, which are mounted and matched in ceramic packages. These are integrated with circuitries on printed circuit boards, resulting in a Doherty amplifier module.

In multi-band transceiver chains, signal filtering is required for each supported radio band. This demands solutions which allow for a closer integration of filters for different bands by enabling volume and weight reduction. Two concepts of resonators with multiple resonant frequencies are key building blocks for multi-band filters. One considers a coaxial resonator with two conductive posts of different size inside, which determine the two resonance frequencies of a dual-frequency resonator. This concept requires a minimum distance between the resonant frequencies as an inherent limitation, which prevents frequency ratios close to 1. The second approach provides full flexibility on the positioning of the resonant frequencies and allows optimizing the volume and resonator quality.

The two cavities are intertwined while the top surface of the lower one determines the lower surface, as an enlarged post of the upper one. The theoretical and experimental evaluation of these dual-band filter solutions, carried out for frequency bands of 2.6 and 3.5 GHz, prove their feasibility and demonstrates a volume reduction of up to 40 % compared to the volume of two single band filters taken individually and separated.

These two concepts of dual-band frequency resonators allow to be combined, resulting in a three-band resonator solution, anticipating significant benefit on volume and weight reduction. The presented concepts for multi-band filters show a variety of possibilities on the port implementations. Coaxial connectors at the input and output can be assigned to one, two or even three bands allowing a flexible adaptation to the transceiver architecture, especially to the number of amplifiers or antennas used to cover the envisaged radio bands.

The list of key building blocks for multi-band transceivers has to be completed by the antennas. Many activities are focusing on broadband solutions, but significant research includes also multi-band approaches at frequencies up to 6 GHz and covering radio bands which are further apart. Antenna topologies based on substrate integrated waveguide (SIW) technology [7][8], which allows for more than 1 GHz bandwidth, were proved for the range of 3.5 to 4.5 GHz and provide advantages on the integration environment. The shielded feeding line prevents coupling with adjacent components, and parasitic surface waves are minimized on the radiating element, which is realized as a slot in a conductor sheet. This eases a compact integration of the antenna in the system. Multiband solutions ask also for different resonant structures or mechanisms realized as a radiating patch on a printed circuit board for the lower frequencies, and a radiating slot within the patch serving the upper frequencies. This allows a free positioning of the two radiation frequencies, but shows a restricted relative bandwidth around 1.5 % to 3 %, as evaluated for radio bands at 2.6 and 3.5 GHz.

The building blocks for multi-band transceivers promote carrier aggregation in radio frontends for increased capacity of mobile systems, as for instance touched in Section 6.5. This is supported by the expected technology evolution on RF data converters with high sampling rates and decreased power consumption and by decreased fabrication costs of GaN based amplifiers, when moving from silicon carbide (SiC) to silicon (Si) as substrate material, which allows the fabrication on larger wafers.

16.2.3 Multi-antenna Transceivers

In this section, concepts and solutions for multi-antenna transceivers are presented, which take into account in which frequency range they are going to operate. Section 16.2.3.1 considers the sub-6 GHz region, whereas Section 16.2.3.2 deals with the mmWave domain.

16.2.3.1 Solutions for Base Stations at Lower Frequencies

Approaches for compact multi-antenna system implementation address multiple signal generation and the chains to connect the antenna elements, including amplification and filtering and resulting in a complete transmitter setup.

A compact solution for RF signal generation is achieved by integrating several transmit chains in a single field programmable gate array (FPGA) or application-specific integrated circuit (ASIC) as digital transmitters. Each chain comprises the generation of a binary pulse train, which includes the modulated RF waveform positioned at the targeted carrier frequency. The signal encoding is based on pulse-width modulation combined with delta-sigma modulation at sampling rates around 24.5 Gsamples/s, enabling high coding efficiency and a large spurious-free bandwidth [9]. The concept is

proved with 8 chains integrated in an FPGA, whereby the number of chains is limited by the high-speed interfaces realized by the board implementation. It allows to position carriers of 5 or 10 MHz bandwidth between 0.9 and 4 GHz, meets the required signal performance, and provides a coding efficiency of 50 % and 100 MHz of spurious-free bandwidth. Within this bandwidth, the mandatory requirements on spectral emission are met. If a spurious-free bandwidth of two times the radio band is provided, no additional filtering is required for spurious emissions and noise suppression besides the radio band (RB) filtering. These results show significant performance improvements compared to prior research results [9]. Further progression is expected when utilizing new FPGA families with a higher sampling rate and an increased number of high speed interfaces used as RF signal ports.

The RF digital signal provided by the FPGA has to be amplified to achieve the signal level targeted at the antenna input. Switched mode amplifiers allow to maximize efficiency as well as to utilize the digital modulation scheme available at the differential interface. Realized as an integrated circuit, it allows to design amplifiers with a form factor suited for mounting behind the antennas, adapted to the grid of the antenna array and enabling a compact system integration. Prototypes realized in GaN technology provide 38 dBm output power in the frequency band between 3.4 and 4.2 GHz [10]. The amplifier is followed by a filter to suppress the unwanted spurious emissions and noise signals. Figure 16-2 shows a schematic architecture of the discussed multi-antenna transmitter with digital transmitter, PA, filter and antenna as the mandatory parts.

Fibre-to-the-antenna (FTTA) links are a promising solution to connect large amounts of distributed antennas to a central BS. In the transmitter architecture, these links are placed at the RF interface between the central multi-chain component (digital transmitter at RF frequency) and the individual RF chains (amplifier, filter) close to the antennas, as depicted in Figure 16-2. Investigations with vertical-cavity surface-emitting lasers and electro-absorption modulators (EAMs) as electrical-optical converters (E/O), and with photo-diodes and EAMs as optical-electrical converters (O/E), chosen as components of low or acceptable costs and operated with multi-mode fibres, show their potential on high data rate transmission [3].

16.2.3.2 Enablers for mmWave Transceivers

Concepts exploiting the large spectral resources available in mmWave bands ask for different building blocks to enable the envisaged radio interfaces. They have to support transceiver systems which include large antenna arrays for spatial directivity based on different beamforming schemes.

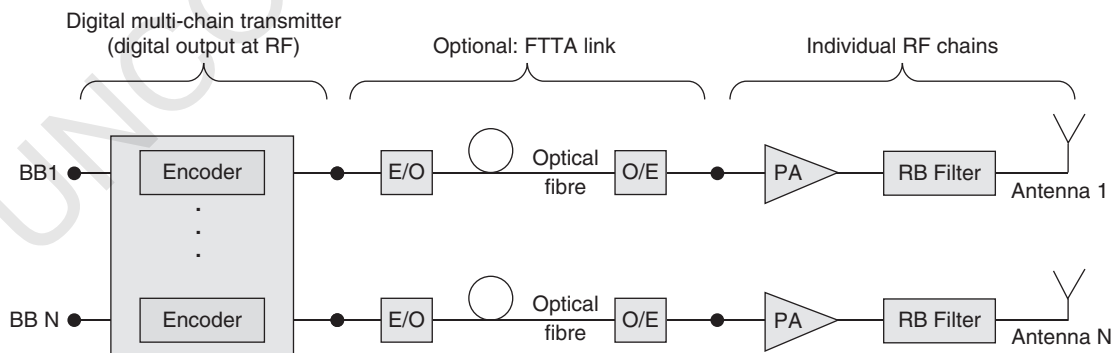


Figure 16-2. Architecture of a possible multi-antenna transmitter.

Antenna arrays combined with large bandwidths are a great challenge for HW implementation, and power consumption will essentially limit the design space. Analogue or hybrid beamforming are currently considered as possible solutions to reduce the power consumption. However, these systems highly depend on the calibration of the analogue components. Another major disadvantage is the large overhead associated with the alignment of transmit and receive beams of the BSs and mobile ends. Specifically, if a high gain is needed, the beamwidth has to be small, thus the acquisition and tracking of the optimal beam alignment in a dynamic environment is very challenging, as discussed also in Section 11.5. Alternatively, power consumption can be reduced by decreasing the resolution of the ADCs, see also Section 11.5.3. This solution has the advantage of keeping the flexibility of fully digital processing, while having reduced power consumption. This is especially advantageous for the case of multiple users transmitting to one access point.

Since hybrid beamforming is the dominating technology used for mmWave prototyping systems today, see also Section 11.5.2, low resolution ADC systems should be compared to it. It was shown in [3], based on a theoretical evaluation and not considering the beam training overhead, that a low-resolution ADC based massive beamforming approach in many situations outperforms hybrid beamforming in terms of throughput and energy efficiency. Overall, this technology might be an interesting new candidate for future mmWave mobile broadband systems.

For providing local oscillator (LO) signals to transceivers for large antenna arrays, one may refrain from using a centralized LO at the high frequency, and instead consider an architecture based on distributed LO generation [2]. There, only the reference signal is common, which is much easier to distribute due to its lower frequency, and the approach has the advantage of phase noise suppression for signals combined at the antenna elements, as the phase noise is then uncorrelated. To support this approach, research was conducted on frequency generation for 28 and 60 GHz carrier frequencies, and it was shown that phase-locked loops (PLLs) with an 8 GHz tuning range are suitable for integration on transceiver chips in 28 nm silicon-on-insulator (SOI) complementary metal oxide semiconductor (CMOS) technology.

The development of amplifiers required for mmWave applications is determined by key performance indicators (KPIs) and parameters such as frequency, bandwidth, efficiency, output power, linearity, gain and cost. To address the latter, the cost-effective nanoscale bulk CMOS technology can be exploited. Its potential was demonstrated by investigations on three power amplifier circuits by playing with the trade-off between output power, bandwidth (BW) and power added efficiency (PAE) [2]. Here, the design was focused on efficiency, yielding 37.6 % PAE combined with 16.8 dBm output power, and 3 GHz BW at 28 GHz operating frequency. An alternative solution provided 6 GHz BW at the same frequency at a similar output power of 16 dBm, but with a reduced PAE of 21 %. With the third solution, the BW was further increased to 28 GHz between 29 and 57 GHz, the PAE maintained at 22%, but the output power was 13.4 dBm only. This shows the potential for designing amplifiers adapted to desired applications.

A further significant building block for multi-antenna systems considers antennas of small form factor and low cost. Planar antennas realized in substrate integrated waveguide technology provide such advantages and ease the system integration due to their inherent shielding of the feeding lines also at mmWave frequencies [2]. Thus, the mounting of the amplifier on the antenna substrate allows a close integration of these building blocks, including a co-optimization of their interconnection with the advantage of reduced reflection of signal power at the antenna port and size reduction of this sub-system.

16.2.4 Full-duplex Transceivers

The challenge of in-band full-duplex (IBFD) transceivers is to simultaneously transmit and receive at the same frequency band. To be able to listen and decode a low-power received signal while transmitting a high-power signal, it is necessary to mitigate the self-interference (SI) signal caused by leakages between the transmitter (Tx) and the receiver (Rx). To avoid placing the computational load of SI cancellation (SIC) on the user equipment (UE), IBFD is considered here only at the BS. The scenario considered is a frequency division duplex (FDD) or time division duplex (TDD) system, where a UE1 transmits to the BS on a certain frequency, while the BS simultaneously transmits to a UE2 on the same frequency. UE1 and UE2 are chosen to be spatially separated to avoid UE1 to interfere towards UE2. Moreover, a small cell scenario is considered, where there is a smaller difference between transmit and received power than in large cell scenarios, thus requiring less isolation between transmitter and receiver, and therefore less SIC capability. Such transmission schemes can bring an important capacity gain compared to standard TDD [11] or FDD.

The SIC is done in several steps, as shown in Figure 16-3. First, a circulator with only one antenna can provide around 20 dB isolation between Tx and Rx ports. Alternatively, two antennas can also be used to further increase the isolation. Then, analogue cancellers (RFSIC) must be used in order to avoid the saturation of the low noise amplifier (LNA), and to allow the ADC to convert the useful signal within the adapted power range. Finally, a digital SIC (DSIC) after the ADC can be applied to remove the remaining part of the SI. The receiver gain must be set in order to not saturate the ADC. An additional hybrid SIC (HSIC) may be also used between the RFSIC and the HSIC. It combines a duplication of the main Tx chain (the so-called auxiliary chain), with a digital linear filter whose role is to take into account the effect of the RFSIC, circulator, coupler, etc.

Figure 16-3 presents the architecture. It is based on a classical 2x2 MIMO transceiver. Only a few functions have been added to manage the swap between TDD MIMO and IBFD SISO.

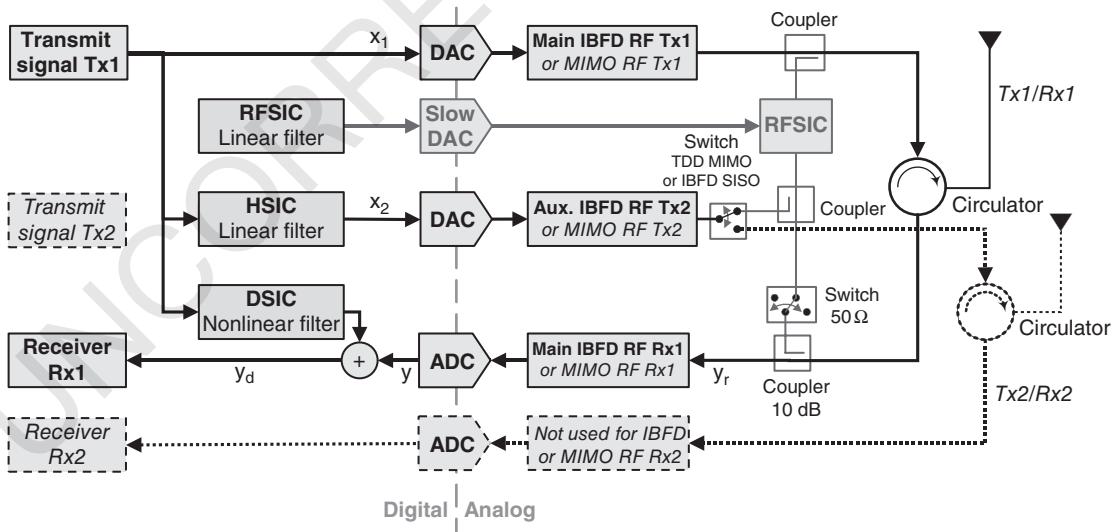


Figure 16-3. Architecture overview of the various stages of the IBFD SIC. The dashed parts are not used in the IBFD implementation.

RFSIC mitigates the strong part of the SI and has low tracking capability. It is based on a digitally controlled vector modulator component. The HSIC mitigates the remaining linear component of the SI. The HSIC can easily manage long delays and provide medium tracking capability. The DSIC cancels the nonlinear contribution of the SI. It is a fully adaptive filter working after the ADC.

Each canceller has been evaluated in a standalone way on real signals thanks to a test bench based on a SDR transceiver board. Globally, with less than 20 dB antenna subsystem isolation, both analogue and digital cancellers together are able to reduce by 85 dB the SI signal over a 40 MHz bandwidth while receiving a useful signal. The main limitation of the current solution is the noise generated by the analogue cancellers. Extra nonlinear distortions of the analogue cancellers are also a limiting constraint. Nevertheless, the depicted solution can lead to 69 % capacity gain on the downlink (DL) on a system level [11].

16.2.5 Techniques for the Enhancement of Power Amplifier Efficiency

As it has been pointed out above, low overall power consumption is a highly desirable feature, in particular because it can increase the battery life of user devices. In this context, poor power added efficiency (PAE, i.e. the ratio between the difference of the RF output power and the RF input power to the DC input power) of the RF PA is one of the primary factors that decrease the battery life. Accordingly, the following two techniques target to enhance the PAE in 5G systems.

16.2.5.1 Envelope Tracking for RF Power Amplifiers

In PAs for 2nd generation (2G) mobile radio, where the output signal has near-constant amplitude, a fixed supply has been used to power the PA achieving high PA efficiency. However, for 3G/4G devices where variable envelope signals are the norm, a fixed supply causes power to be wasted when a device transmits below its maximum output power. Envelope tracking (ET) is a well-known technique to improve PA efficiency in that case [3]. The principle behind ET is to continuously adjust the power supply voltage to follow the envelope of the transmitted signal in order to ensure the desired PA efficiency. This way, ET optimizes PA power consumption. Alternatively, ET can linearize a PA by reducing the amplitude-to-amplitude (AM-AM) distortion, which allows reducing out-of-band emissions or the transmission of higher order modulation schemes and, therefore, higher data rates. The main elements of the ET procedure comprise the detection of the RF magnitude at I/Q sub-sample rate and, accordingly, the update of the power supply voltage. It is of highest importance that this so called “envelope path” is well calibrated to support the targeted waveform and take into account system imperfections such as the delay and gain offsets introduced by the circuit blocks in the signal path. Looking at 5G systems, the main challenges for ET performance compared to LTE/4G include the wider bandwidths (well above 20 MHz), the higher carrier frequency (>6 GHz) and the new waveforms used. A wider signal bandwidth has a direct impact on the susceptibility of the system to delay imbalances; doubling the bandwidth requires the delay mismatch to be halved. A higher signal bandwidth also has impact on the bandwidth of the envelope signal. Moreover, in higher carrier frequencies, PA memory effects become non-negligible and impact out-of-band emissions. In [3], a suitable PA model was adopted to evaluate the memory effects for ET and showed that distortion cannot be counteracted efficiently. Especially at higher frequency offsets there is a mismatch between supply voltage and RF signal magnitude, which results in some higher-order distortion terms. While

nearby spectral emissions are suppressed, further distant spectral emissions actually degrade when compared to a fixed supply voltage. Finally, waveforms such as filtered orthogonal frequency division multiplex (OFDM) can provide better frequency localization than LTE, and the advantage is that interference with other users is reduced due to much lower unwanted emissions; however, once PA nonlinearity is considered, much of the spectral properties are lost due to intermodulation products polluting nearby frequencies. All things considered, it is proved that fully dynamic ET may be challenging for 5G and less aggressive techniques may be preferred, like average power tracking (APT), where the PA supply voltage is adjusted based on the average power over a certain period.

16.2.5.2 Weighted Selective Mapping and Digital Pre-distortion Techniques for 5G Waveforms

5G waveform candidates such as Orthogonal Frequency Division Multiple Access (OFDMA) and filterbank multi-carrier (FBMC) are based on a multi-carrier technique. They are robust to the frequency selective fading channel and can achieve a high data transmission. However, the major drawback of multi-carrier is a high peak-to-average power ratio (PAPR), and this drawback causes signal distortion and high energy consumption [12]. A PA is very sensitive w.r.t. to its operational area, and hence a high PAPR causes a nonlinear operation problem and consequently signal distortion.

A new PAPR reduction technique named Weighted Selective Mapping Technique (WSLM) was developed and its complementary cumulative distribution function (CCDF) performances evaluated in OFDM and FBMC systems [3]. In addition, a memory polynomial digital pre-distortion (DPD) was used to compensate for nonlinear behaviour of a PA. It applies inverse distortion to the input signal of the PA in order to compensate the distortion generated by it. The performance of WSLM and DPD techniques was evaluated and about 2.5 dB gain was achieved under the given test configuration.

16.3 Solutions for Digital HW Implementation

16.3.1 Requirements on 5G Digital HW

Device chipset and platform suppliers will face the need to support an increasing number of RATs together with the trend toward a highly diverse set of device form factors, with cars, wearable and machine-type devices joining tablets and smartphones. This progressive increase in form factor diversity is driving multiple levels of platform support and capability, making transceiver complexity a key challenge for 5G devices. In addition, 5G devices are expected to integrate and interact with a multiplicity of sensors (e.g., those related to location and positioning, environmental conditions, image processing etc.) that will provide context awareness to the communication and the deployed RAT, which will allow improvements in the efficiency of existing services, and help to provide new and more user-centric and personalized services.

A similar heterogeneous situation to that of 5G devices is also valid for 5G network elements. Since cost and flexibility of deployment will also be important factors, a shift towards SW-based implementations and virtualization technologies will be required, as detailed in Section 5.2.3. Ideally, these SW-based implementations should be as HW-agnostic as possible and, at the same time, should be supported by versatile, reconfigurable, and flexible HW platforms that are able to cope with all the functionalities needed and invoked from the SW domain.

16.3.2 Complexity Analysis of the Individual Implementation of New Waveforms

An indicative way to indirectly quantify the implementation cost of digital HW is the analysis of the computation complexity of digital baseband processing blocks. The waveform generation and recovery are two of the basic digital signal processing blocks that need to be flexible and scalable in 5G system architectures due to the various possible use cases, scenarios and their requirements. In MIMO systems, for example, one waveform generation and recovery block is necessary for each antenna, and for carrier aggregation the same is true for each component carrier. In addition to that, if multiple subcarrier spacings are employed for different time-frequency resource blocks, separate waveform generation and recovery may need to be implemented for each of them.

In this subsection, the computational complexity of the waveforms proposed for 5G and beyond are analysed by considering individual implementations, and in the next subsection, a harmonized implementation of multiple waveforms as proposed in Section 11.4.4.

Multi-carrier (MC) and OFDM-based single carrier (SC) modulations considered for 5G and beyond include:

- Conventional cyclic prefix OFDM (CP-OFDM) with windowing [13].
- Discrete Fourier transform (DFT)-spread-OFDM, also referred to as DFT-s-OFDM.
- Single carrier frequency division multiple access (SC-FDMA).
- Zero tail DFT-s-OFDM (ZT-DFT-s-OFDM) [14].
- FBMC with offset quadrature amplitude modulation (FBMC-OQAM) [15] and with quadrature amplitude modulation (FBMC-QAM) [16].
- Filtered multitone (FMT), a.k.a. pulse-shaped OFDM (P-OFDM) [17].
- Universal filtered multicarrier (UFMC or UF-OFDM).
- Cyclic convolution based FBMC, a.k.a. generalized frequency division multiplexing (GFDM).

In this section, the complexity is quantified and compared in terms of the total number of real multiplications and additions.

We consider the signal processing operations involved in the generation of the MC and SC signals, as well as the recovery of the subcarrier/subchannel/subband signals and equalization in the presence of multipath propagation. Here, we do not consider the operations involved in channel estimation or calculation of the equalizer coefficients. The first reason is because those signal processing tasks are not in the user data chain, which is the one that concentrates the processing burden, and the second is because of the many existing algorithms for those tasks making the choice of one not trivial. Moreover, we assume that all systems are perfectly synchronized.

Because all waveforms are based on MC modulation, we assume that a total of N subcarriers are available out of which N_f are occupied with symbols. We will consider first the number of real-valued multiplications and additions to transmit one block of N_f symbols.

Since the transmitter (Tx) and receiver (Rx) of a CP-OFDM system are basically built with one single inverse fast Fourier transform (IFFT) and fast Fourier transform (FFT) and, possibly, a windowing operation, the SC waveforms cyclic-prefix-based DFT-s-OFDM and ZT-DFT-s-OFDM have the same complexity of CP-OFDM plus the DFT spreading part, which slightly increases the complexity. In the case of ZT-DFT-s-OFDM, the complexity is similar to cyclic-prefix-based DFT-s-OFDM, but the two main differences are that no cyclic prefix is attached and zero valued headers and tails in a subchannel level are added.

One of the basic and common building blocks for all waveforms is the FFT/IFFT. The number of real multiplications and additions of an M -point FFT/IFFT using a split-radix algorithm are given by [18]:

$$\begin{aligned} C_{m_{FFT}}(N) &= N(\log_2(N) - 3) + 4, \\ C_{a_{FFT}}(N) &= 3N(\log_2(N) - 1) + 4. \end{aligned} \quad (16-1)$$

Assuming an FBMC-OQAM system where the prototype filter has length KM , two approaches can be adopted for the generation and recovery of the MC signal: the polyphase-based and the frequency-spread based structure.

We consider first the complexity of FBMC-OQAM implemented with a structure based on the polyphase decomposition of the prototype filter and using a direct form realization of the polyphase components (PPC) [18]. The Tx is composed of 3 steps after the OQAM modulation:

- Phase rotations to get linear phase filters in each subcarrier, i.e., $4Nf$ real multiplications.
- IFFT.
- Polyphase filtering followed by block overlapping of 50%, i.e. $4KN$ real multiplications.

At the receiver (Rx) side similar operations in the inverted order are implemented including one more step: polyphase filtering, FFT, multitap channel equalization per sub-carrier with an equalizer of length Leq , resulting in a complexity of $4NfLeq$ and the OQAM demodulation. The phase rotations at the Rx side can be embedded in the equalizer coefficients.

The second approach is a frequency domain filtering, also known as frequency spread based (FS)-FBMC-OQAM [19], featuring also a prototype of length KN and designed using the frequency sampling approach [20] with only $2(K-1)$ non-zero coefficients in the frequency domain. In this case, the structure changes drastically. The subcarrier signals have to be spread over $K+1$ frequency domain samples, and each of them multiplied by one of the prototype frequency domain coefficients. The complexity of the frequency domain filtering encompasses $8N_f(K-1)$ multiplications. The overlapping parts in frequency domain are all added and then transformed with an IFFT of size KN . Finally, the first $M/2$ samples of a given block of KN IFFT outputs are added to the last $M/2$ samples of the previous block to generate the serialized time domain signal. At the Rx side, the inverse operations are done. In addition to the FFT at the Rx, a frequency domain filtering with $16N_f(K-1)$ multiplications is also performed.

The FMT/P-OFDM waveforms are similar to the FBMC/OQAM case, but there is neither OQAM nor a 2-stage up-sampling. Nevertheless, for the calculation of the number of operations per sample, we need to consider the lower sampling rate.

The UFMC system can be parametrized between two extremes: on one end, the whole CP-OFDM signal is filtered by one filter to reduce the out-of-band radiation. At the other end, each or a minimum number of resource blocks is transformed with the IFFT and filtered with its own filter. In an UFMC system with maximum granularity, NB resource blocks each with M subcarriers require NB FFTs of size N , where each of them has only NNB non-zeros inputs.

The modulation is performed in the following steps: First, the signal of each subband is spread over the whole symbol length and transformed into the frequency domain. Then, the filtering is performed in the frequency domain, and the sum of all subbands is converted into the time domain [21].

Instead of filtering and then transforming, a non-matched filtering is applied in the frequency domain [22]. The Rx then has 3 steps:

- Windowing in the time domain
- FFT transformation of size $N_u M$ with zero padding and half of the outputs thrown away
- Frequency domain filtering and equalization.

The GFDM modulation scheme is based on circular convolving each subcarrier in a block of data with a filter kernel. In contrast to OFDM, a cyclic prefix is added per block and not per symbol [23][24]. Since a circular convolution can be calculated as a multiplication of two vectors in the frequency domain, then both transmitter and receiver can be efficiently implemented using the FFT.

Out of a total number of subcarriers N only N_f are used. MB symbols per subcarrier are combined to form one transmission block. In total $N_f MB$ data symbols can be transmitted per block. The prototype filter is designed to overlap with N_a adjacent subcarriers and it is typically chosen to be an RRC filter $N_a = 2$.

As described in [23], excluding the trivial operations like reordering, the following signal processing tasks need to be performed at the transceiver:

- Transformation of the data signal of each subcarrier into the frequency domain
- Filtering in the frequency domain
- Transformation of the signal into the time domain

The details of the corresponding receiver are described in [24]. It is important to mention that since the subcarriers are overlapping it is necessary to cancel this interference to achieve a sufficient performance. In [24], the authors use the detected symbols to subtract the interference to adjacent subcarriers in an iterative fashion. For a constellation as large as 64QAM it was shown that $J = 8$ iterations are sufficient. The receiver can be divided into the following signal processing tasks:

- Transformation of the signal into the frequency domain
- Channel equalization
- Filtering in the frequency domain
- Iterative interference cancelation

Let us now consider a numerical evaluation of the individual waveform implementations. For that, we calculate here only the complexity at the base station. We utilize the complexity formulas presented in detail in [25]. The metric utilized is the number of multiplications and additions normalized by the number of QAM symbols transmitted across the used subcarriers. We assume a similar overhead in terms of training or reference signals for all waveforms.

Moreover, we consider here four 3GPP New Radio (NR) wideband parameters [26]:

- 1) NR downlink with BS wideband and UE narrowband allocation (UE: 1 MHz, BS: 100 MHz)
- 2) NR downlink with wideband allocation (100 MHz)
- 3) NR uplink with BS wideband and UE narrowband allocation (UE: 1 MHz BS: 100 MHz)
- 4) NR uplink with wideband allocation (100 MHz)

In scenarios 1 and 3, six resource blocks are allocated for each UE, and 85 mobiles are served simultaneously. In scenarios 2 and 4, only one UE is served, and all 6600 available subcarriers are allocated to it.

The parameters have been chosen as given in Table 16-1.

Table 16-1. Parameters for the numerical complexity analysis.

• CP-OFDM	<ul style="list-style-type: none"> • FFT size: $N = 8192$ • Number of active subcarriers: $N_f = 6600$ (wideband), 6120 (narrowband MS, wideband BS) • Cyclic prefix length: $L_{CP} = 576$ • Number of RB: $N_{min}^{RB} = 6$ (narrowband) or $N_{max}^{RB} = 550$ (wideband)
• DFT-S-OFDM/ • ZT-DFT-s-OFDM	<ul style="list-style-type: none"> • Size of the resource blocks: $M = 12$ • Size of the small FFTs: $MN_{min}^{RB} = 72$ (narrowband) or $MN_{max}^{RB} = 6600$ (wideband)
• FBMC	<ul style="list-style-type: none"> • Time domain overlapping factor $K = 4$ • Number of equalizer taps/subcarrier: $L_{eq} = 3$ (polyphase) • Frequency sampling: $L_{eq} = K$
• UFMC	<ul style="list-style-type: none"> • Number of subcarriers/RB: $M = 12$ • Filter length: $L = L_{CP} + 1$ • Frequency oversampling factor: $N_u = 2$ • Size of narrowband (NB) FFT: $N_{NB} = 64$
• GFDM	<ul style="list-style-type: none"> • Number of symbols/subcarrier: $M_B = 4$ • Number of Overlap subcarriers: $N_a = 2$ • Number of SIC iterations: $J = 8$

In Figure 16-4, the BS complexity results related to the different waveforms for the different scenarios are shown.

16.3.3 Complexity Analysis of a Multi-waveform Harmonized Implementation

A flexible implementation integrating multiple waveforms in a single harmonized solution can be very useful to select the waveform that better matches a particular communication scenario and to reduce implementation costs. For this particular purpose, a generic MC waveform implementation was presented in Section 11.4.4. In such a tuneable waveform implementation, by selectively enabling or disabling particular blocks, one out of six different MC waveforms could be generated. The waveforms of interest include classical CP-OFDM, windowed OFDM (W-OFDM), P-OFDM and SC-FDMA or ZT-DFTs-OFDM. Furthermore, the framework is also valid to represent waveforms of the FBMC family such as FBMC-QAM and FBMC-OQAM.

Firstly, focusing on the transmitters for the independent waveforms, it can be noted that they all include the following blocks: one for serial-to-parallel data conversion, one for QAM constellation mapping, a subcarrier mapping and pilot insertion block, a block for MC modulation through FFT and, before sending the data to the channel, one parallel-to-serial data conversion block. However, recall that there are some extra modules needed for specific waveforms, which should be considered within the harmonized implementation:

- DFT spreading/de-spreading: This block is intended to perform the spreading/de-spreading operations necessary for the ZT-DFT-s-OFDM waveform transmission/reception. It has the same cost as an FFT of size M .

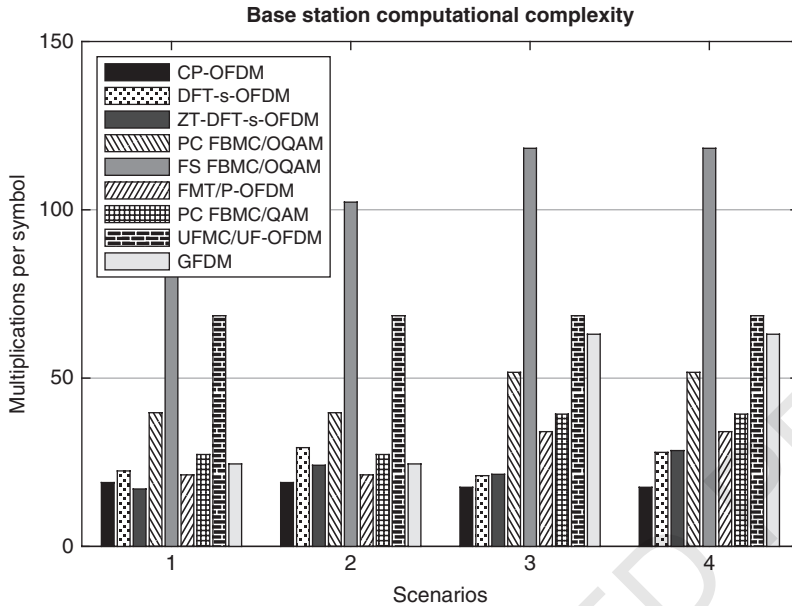


Figure 16-4. BS computational complexity in terms of number of real-valued multiplications per Tx/Rx symbol – NR wideband.

- **OQAM pre-processing/post-processing:** The FBMC-OQAM waveform needs a specific module for complex to real conversion of QAM symbols, up-sampling, and time staggering. At the receiver, these operations obviously need to be reverted.
- **Reconstructing IFFT block:** Because of its importance for reducing the complexity of FBMC-OQAM, we propose the simultaneous calculation of the FFTs of size N for the two real-valued inputs involved, via a single FFT of size N with complex-valued inputs, as shown in [27]. After performing the IFFT, the additional total cost of reconstructing each individual FFT for both two real-valued inputs is $8N$ multiplications and $4N$ additions.
- **Polyphase network (PPN):** The implementations of all the waveforms that include filtering or windowing in the time domain (W-OFDM, P-OFDM, FBMC-QAM and FBMC-OQAM) include a PPN, where the input data is convolved with a filter. For the complexity evaluation in the next section, it will be assumed that the prototype filter used in the PPN has real coefficients, i.e. it is symmetric in the frequency domain, the most common assumption in waveform implementations.

16.3.3.1 Complexity Evaluation

The harmonized implementation at the minimum complexity requires: one FFT block of size N (common to all waveforms); one block to rebuild the FFT of two real-valued inputs (required for FBMC-OQAM); two blocks of real-valued PPN filtering; and one block for DFT spreading of size M , necessary for ZT-DFT-s-OFDM. Considering the complexity analysis carried out in the previous

section for new waveforms, the aggregated complexity *cost* in terms of multiplications, additions, and floating point operations per second (flops) is [28]

$$\begin{aligned} C_{m_{\text{HARM}}}(N, K) &= N \log_2 N + 4NK + 5N + M \log_2 M - 3M + 8, \\ C_{a_{\text{HARM}}}(N, K) &= 3N \log_2 N + 4NK - 3N + 3M \log_2 M - 3M + 8, \\ C_{f_{\text{HARM}}}(N, K) &= 4N \log_2 N + 8KN + 2N + 4M \log_2 M - 6M + 16. \end{aligned} \quad (16-2)$$

The complexity cost of a solution where all the waveforms are drawn together as standalone implementations is found by adding the number of multiplications, additions, or flops of all of them. The total result is

$$\begin{aligned} C_{m_{\text{NOHARM}}}(N, K) &= 6N \log_2 N + 8NK - 8N + M \log_2 M - 3M + 28, \\ C_{a_{\text{NOHARM}}}(N, K) &= 18N \log_2 N + 8NK - 22N + 3M \log_2 M - 3M + 28, \\ C_{f_{\text{NOHARM}}}(N, K) &= 24N \log_2 N + 16NK - 30N + 4M \log_2 M - 6M + 56. \end{aligned} \quad (16-3)$$

Assuming the typical values $K=4$ and $M=12$ and a number of subcarriers ranging from 16 to 4096, Figure 16-5 and Figure 16-6 show the complexity in terms of multiplications, additions and flops of the harmonized and non-harmonized implementations for different values of N . From basic calculations based on the values in the figures, it can be observed that the typical savings due to harmonized implementation range between 60–75%.

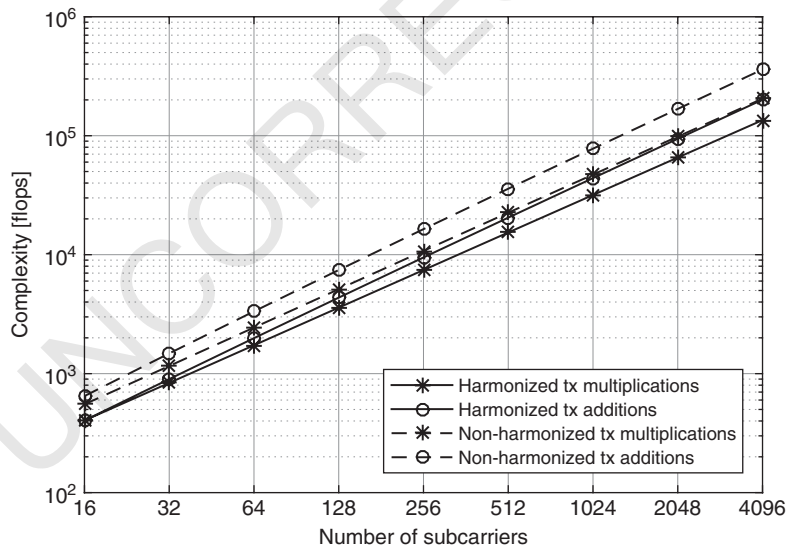


Figure 16-5. Multiplications and additions of the proposed harmonized and the non-harmonized implementation of the six waveforms.

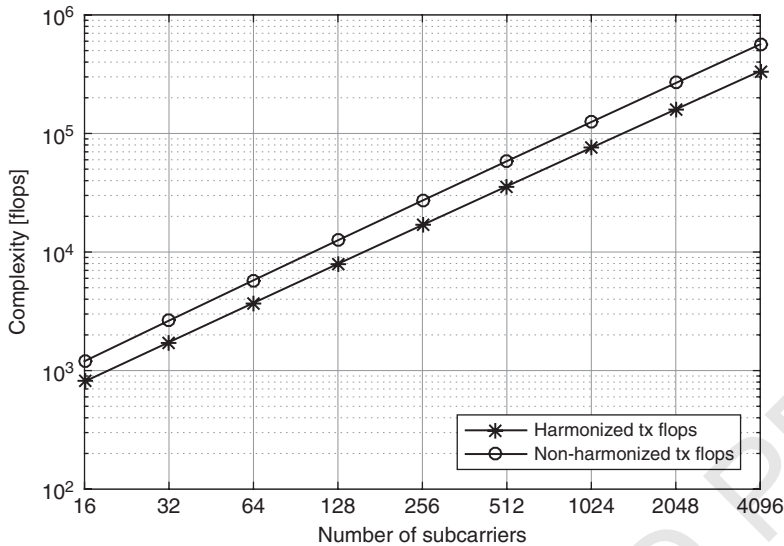
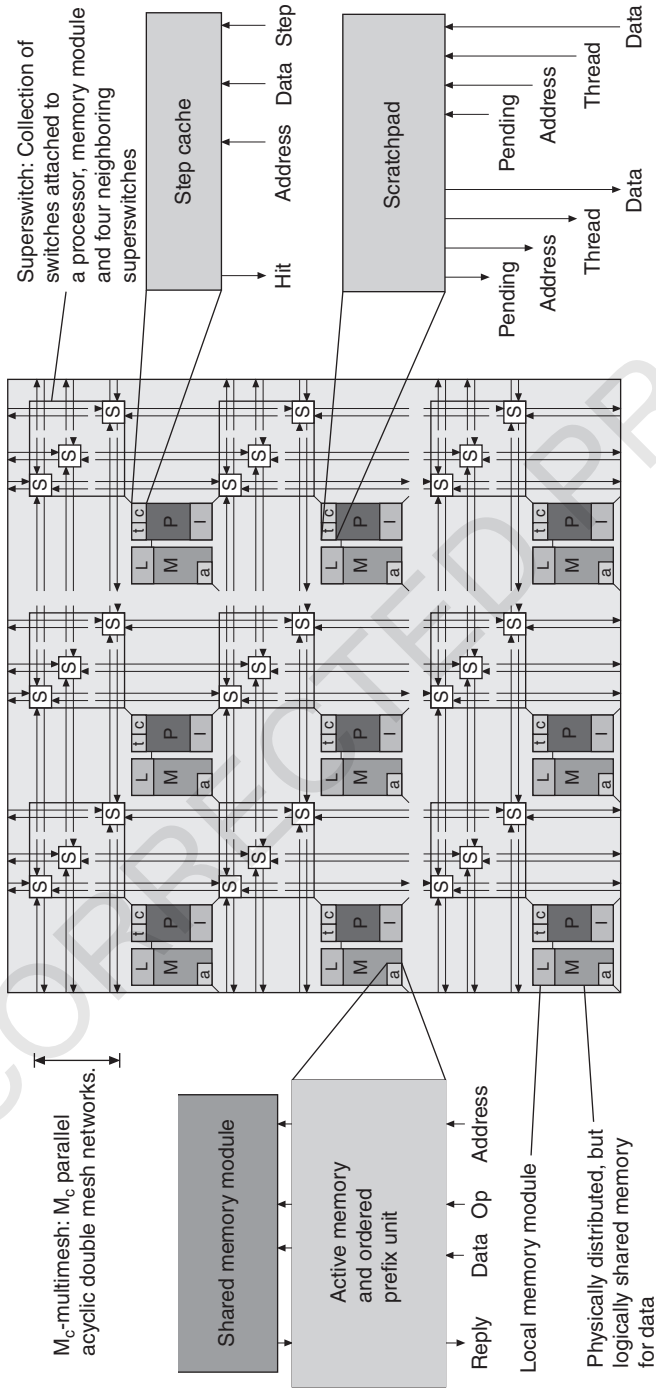


Figure 16-6. Number of flops of the proposed harmonized and the non-harmonized implementation of the six waveforms.

16.3.3.2 Multi-processor Baseband Architectures for 5G Network Elements

Current baseband processing architectures, such as the ones required to implement the waveforms whose complexity has been analysed in the previous sections, typically rely on a combination of general purpose processor, DSP processors and dedicated HW due to stringent performance and power-efficiency requirements. The flexibility and programmability of these solutions featuring fixed partitioning is typically limited since the DSPs are not efficient in general purpose processing. Further, general purpose processors are not efficient in dedicated tasks, and it is impossible to use HW accelerators for anything other than their dedicated tasks. While it is not possible to entirely remove these limitations, with a smart enough architecture it is possible to have fully programmable and flexible computing solution for 5G baseband computation, which can be suitable for, implementation in network elements, for example.

REPLICA is an architectural framework (or family of chip multi-processors) aimed to solve the performance and programmability bottlenecks of current multi-core processors [29]. The performance comes from a scalable latency hiding (a technique to eliminate delays caused by the shared memory system) combined with a low-cost inter-thread synchronization mechanism and efficient exploitation of low-level parallelism. Programmability is achieved via support for architecture-independent abstraction, strict memory consistency and synchronous, more deterministic execution than in current alternatives. A REPLICA chip multi-processor (CMP) consists of P processors, MS shared memory modules, MI local memory modules, P instruction memory modules and a high-bandwidth network for connecting the elements, as shown in Figure 16-7. REPLICA processors are Tp -way multi-threaded to support latency hiding of the shared memory system, where Tp is a design-time parameter chosen so that the latency of typical memory accesses gets hidden.



M_c -multimesh: M_c parallel acyclic double mesh networks.

Superswitch: Collection of switches attached to a processor, memory module and four neighboring superswitches

Figure 16-7. REPLICIA CMP [29] with P processors, M_c -way multi-mesh network and P active memory modules (P=processor core, I=instruction memory module, t=scratchpad, c=step cache, a=active memory unit, M=shared memory module, L=local memory module and S=switch).

They feature a chained very long instruction word organization of functional units to support execution of dependent sub-instructions within an instruction and have $T_p + 1$ -stages pipeline to avoid pipeline delays that affect current processors [30].

Applying the REPLICIA architecture framework for 5G baseband processing requires architectural improvements to fulfil performance requirements with affordable energy efficiency. Modifications to the threading system have been introduced, and an architectural unit to speed up the FFT computation has been added. In terms of the silicon area and power consumption, this solution looks affordable with respect to commercial general purpose processor alternatives [25].

16.3.4 Channel Decoder Implementations for 5G

5G systems should also consider advanced forward error correction (FEC) technologies such as advanced Turbo-codes, low-density parity check (LDPC) codes and polar codes. From the decoder standpoint, such FEC codes are often decoded using iterative decoding strategies, relying on maximum a posteriori (MAP) or similar algorithms. 3GPP has moved forward on the selection of the coding scheme for eMBB, and LDPC and Polar codes are most likely to be considered for the data and control signals, respectively. As the data rates in 5G will be very high also on the terminal side, there is a need to develop high speed channel decoders that are also energy and area efficient. However, for backward compatibility with LTE, it is also of high interest to investigate flexible decoder architectures able to address Polar, LDPC or Turbo codes.

16.3.4.1 Message-Passing Decoding for LDPC Codes

LDPC codes are a class of error correction codes known to closely approach the Shannon limit under iterative message-passing (MP) decoding algorithms [31]. MP architectures are composed of processing units that perform the desired computations and pass messages to each other. The way such architecture applies to LDPC decoding is closely related to the bipartite graph representation of LDPC codes [32]. It comprises two types of nodes, known as variable-nodes (VNs) and check-nodes (CNs), corresponding respectively to coded bits and parity-check equations. Accordingly, an LDPC decoder comprises two types of processing units, namely VNUs and CNU, which exchange messages according to the structure of the bipartite graph.

Two layered decoder architectures for Quasi-Cyclic (QC)-LDPC codes have been designed targeting high data throughput and an efficient use of HW resources. High throughput is achieved by either pipelining the datapath or increasing the HW parallelism in the architecture. Both architecture variants may accommodate two different decoding kernels. First decoding kernel corresponds to the conventional Min-Sum (MS) decoding, while the second corresponds to the Non-Surjective Finite Alphabet Iterative Decoder (NS-FAID), which has been proposed in [33] and it was later shown that NS-FAIDs can provide different trade-offs between HW complexity and decoding performance for both regular and irregular LDPC codes [25].

The layered decoder architectures with both MS and NS-FAID decoding kernels were evaluated in terms of throughput and resource consumption, for both regular and irregular QC-LDPC codes. Implementation results of the proposed architecture for both FPGA and ASIC platforms show improvements in throughput compared to state-of-the-art implementations. For the ASIC implementation of 65 nm CMOS technology, an area-normalized useful throughput of 2293 Mbps/mm² was achieved while the comparable state-of-the-art implementations achieve performance between 106 Mbps/mm² and 1257 Mbps/mm² [25]. It has also been shown

that NS-FAIDs allow significant improvements in terms of both throughput and HW resource consumption, as compared to the conventional MS solution, while also improving the error correction performance.

16.3.4.2 Turbo Decoder Design Optimized for mMTC

For backward compatibility with LTE, it is of high interest to investigate flexible decoder structures able to address Polar, LDPC or Turbo codes. The focus here is on iterative decoders based on MAP or related measures such as max-log MAP. The latter is an approximate low-complexity algorithm with some degradation in performance when compared to a log-MAP or MAP algorithm.

In order to meet very high data rates, typical implementations consider parallelized architectures, in which several MAP instances are used simultaneously to speed up the overall processing. On the other hand, having several parallel instances of such modules increases both the size (and thus cost) and the power consumption of the solution. Such approach could, therefore, be challenged when addressing low-end devices (e.g. connected objects in a Smart Cities use case), for which cost and power efficiency become key characteristics. A flexible turbo decoder has been designed that is able to adapt dynamically to the different services expected for 5G [25]. By configuring the number of active MAP engines, the number of iterations and number of warm-up stages, it is possible to configure the decoder to fulfil latency, throughput, power consumption and performance requirements as needed.

16.4 Flexible HW/SW Partitioning Solutions for 5G

Mobile wireless systems in 5G are envisioned to offer increased performance on top of flexible heterogeneous devices while reducing the overall energy consumption by efficient virtualization and coordination technologies. In order to reduce the number of physical network elements, new systems will need to incorporate virtualization mechanisms, which will be able to cope with the increased network demands without affecting timing constraints and communication overhead. New implementation techniques must be introduced in order to further increase reusability, flexibility and performance along with a reduction of energy consumption at the same time. Some solutions that introduce static and customized functions virtualization able to fully virtualize a network element and offer significant power reduction are already available. In the case of RAN network nodes, existing virtualization technologies perform the entire digital baseband processing part in SW platforms and utilize SDR platforms connected to remote radio heads (RRH) for amplification and transmission purposes, as for instance discussed in Section 6.7. Some implementations are also able to simulate full LTE networks. The most notable among them are the LENA ns-3 [34], which provides full SW implementation of virtual LTE networks, the OpenAirInterface [35], OpenLTE [36] and srsLTE [37].

16.4.1 Architecture for Supporting MAC/PHY Cross-layer Reconfiguration

In this section, we describe an architecture for wireless terminals able to support advanced Medium Access Control (MAC) and physical layer (PHY) reconfigurations. The design has been motivated by the need to improve the terminal capabilities for exploiting context information and adapting the PHY to the application requirements and network operating conditions. Following

this need, PHY-layer improvements can be divided into two main groups: i) those aiming at supporting new PHY primitives, devised to configure the central frequency and the bandwidth of the transceiver, in order to optimize the spectrum utilization and minimize the interference generated to other coexisting links; ii) those aiming at supporting new PHY measurements, in order to characterize the network environment.

The terminal is based on an extension of the so-called wireless MAC processor architecture [25], according to which the terminal transceiver is driven by a programmable state machine which specifies the MAC protocol logic and the configuration of the transceiver and receiver processing chain. The prototype has been implemented on the WARP v3 research platform, by exploiting the implementation of a complete Institute of Electrical and Electronics Engineers (IEEE) 802.11 MAC/PHY legacy stack, already available for this platform. For this reason, the design has been organized in terms of incremental extensions of the IEEE 802.11 implementation. While the generic executor of MAC/PHY state machines was implemented at the firmware level with minimal modifications of the original firmware (devised to take into account the possibility to specify chains of multiple actions), the implementation of advanced PHY layer primitives required to work on the IP cores and HW blocks of the platform [38].

The developed architecture has been applied to evaluate the benefits of dynamic bandwidth adaptation and innovative signalling mechanisms based on tones. Bandwidth adaptation is more powerful than simple adaptive modulation schemes, because it can increase robustness by identifying the most suitable spectrum portion to be utilized for transmissions. Adaptive modulation can still be utilized on top of bandwidth adaptation. Although the idea of bandwidth-dynamic adaptation for wireless links is not new and especially relevant in the cognitive network scenario, the proposed solution is promising, since it does not require an out-of-band or in-band signalling channel to be used for negotiating the bandwidth to be utilized.

16.4.2 Cognitive Dynamic HW/SW Partitioning Algorithm

The static HW/SW partitioning methods may generally lead to reduced energy consumption, but they are not reusable and not reconfigurable thus limiting network upgrades and resource allocation. Also, the processing power cannot be shared among nodes, which can lead to load unbalancing and, thus, a decrease in efficiency. Full virtualization is not always available as the devices of the underlying network might not be able of performing such a task, or virtualization might be limited according to available physical and computational resources. Within this context, the cognitive and dynamic HW/SW partitioning algorithm's task is to provide the best HW/SW partitioning of the 5G network functions for both device and network elements. The algorithm takes the required KPIs for a given scenario and the available HW/SW resources and optimizes for high flexibility, configurability, performance and/or energy efficiency [38].

In [38], the cognitive and dynamic HW/SW partitioning algorithm has been tested considering a dynamic hotspot use case, resulting in the LTE enhanced Node-B (eNB) PHY layer reconfiguration according to pre-specified measured KPIs and optimization goals that lead to higher flexibility, lower execution time and energy consumption reduction. The cognitive and dynamic HW/SW partitioning offers extended flexibility and re-configurability inside a network node, while also ensuring the KPI requirements and further applying an optimized functional partitioning towards user-defined KPIs. In terms of reconfiguration, the proposed solution performs its optimization operation in less than 17 ms for common scenarios.

16.5 Implementation of SW Platforms

For 5G systems, it is of utmost importance that the control and the management of heterogeneous HW infrastructure and devices expected are done in a way that guarantees: (i) an effective (i.e., rapid, easy and dependable) service development and deployment, (ii) the ability to adapt to very demanding and changing contexts of operation, and (iii) to guarantee Quality of Service (QoS) and Quality of Experience (QoE). To ease this burden while efficiently making use of the available resources, it is required the use of two enabling technologies: on one hand the deployment of scalable, flexible and multi-RAT SW networks, and on the other hand, introducing a HW-agnostic operation through the use of programmability and dynamic reconfiguration via interface abstractions/uniform Application Programming Interfaces (API). With the above, changes in context (as for instance users' demands, interference, availability of resources, etc.) as identified by sensors will trigger changes in the operation of radio technologies, but at an affordable control and management overhead.

In the last years, the need to introduce this reconfigurability in the wireless networks has been identified as a key requirement to efficiently deploy and maintain high-performance wireless deployments [39], where many end devices can be densely deployed and may be equipped with different radio communication technologies (Wi-Fi, ZigBee, Bluetooth, 2G/3G/4G...) to cope with different application requirements in terms of data rate, latency, reliability and energy consumption [40]. Different wireless technologies will compete for the same limited spectral resources, and it is therefore very important to deploy a suitable protocol stack with proper configuration settings that adapts to changing wireless and application contexts. Reconfiguration involves the ability to change protocol parameters, a network protocol, e.g., the MAC layer, the radio operation mode (e.g., the modulation and coding scheme, channel frequency, transmitted power, etc.) [41][42].

In order to support the above vision, the selection of a suitable protocol stack and the corresponding configuration parameters should happen in an autonomous way without the involvement of radio or network experts, otherwise the complexity of the network precludes its scalability. In the area of cognitive radio networks, a lot of attention is paid to programmable SW communications architecture [43][44][45][46][47][48][49]. These architectures focus either on a single radio technology or sophisticated SDR platforms. Single radio platforms are cost efficient, but offer limited radio flexibility. SDR platforms, on the contrary, are very powerful, but expensive and consume a lot of energy. Both platforms are expected to coexist in future dense wireless scenarios, although their difference in complexity will result in different levels of adoption. Furthermore, while these platforms provide the ability to adapt the operation of the wireless interfaces to the estimated condition, they do not provide a complete wireless architecture including all required functionalities (apart from re-configuration). In what follows, we describe the functionality that is envisioned to leverage on this re-configurability to optimize performance in 5G systems.

16.5.1 Functional Modules

Versatile wireless interfaces is a key requirement future 5G systems have, as this functionality enables adapting to the estimated context conditions. However, to take advantage of this versatility, the interfaces have to be provided with the adequate functionality to properly estimate the conditions, and to react accordingly in the most appropriate manner, which may involve altering the operation of other wireless interfaces. To satisfy the requirements discussed above in 5G systems, the architecture of future wireless interfaces have to be designed building on three key concepts, namely, flexibility,

reconfigurability and monitoring. This required functionality can be divided into three different areas: the operation of a single device, i.e. intra-node enhancements, the coordination of multiple devices, i.e. inter-node functionality, and the control and optimization of the network, i.e. intelligent programs to optimize performance.

For the single node operation, the most relevant concept is flexibility: to cope with the high variability of 5G application requirements and network topologies, 5G technologies will support advanced reconfiguration capabilities at both the PHY and MAC levels. Architectures are needed, which move from the traditional approach of “one-size-fits-all” MAC/PHY protocol stack to an innovative paradigm of on-the-fly configuration of context-specific stacks, see also Section 6.4, implementing abstracted versions of wireless primitives that sit between the current ossified stacks and the fully-programmable SDR solutions.

The control and optimization of the network deals with the optimization of network performance, and should be based on an information-centric operation and exploit the re-configurability and monitoring features. This requires functionality such as a monitoring library that provides access to the data collected by sensors and monitoring agents throughout the network, which should feed two modules that extract network-wide performance and context estimations to trigger different types of optimizations. For instance, performance degradations or context variations may trigger the activation of additional network elements or switching to a Radio Access Technique that is more robust. Moreover, we also envision the need of a global scheduler, coordinating the operation of the BSs, enabling the C-RAN and vRAN vision as described next, and a service scheduler that coordinates the optimization of all these elements, to enable a smooth operation that precludes conflicts.

Finally, to connect these two areas, a third one is needed: multi-node coordination. Here, we envision functionalities such as: a monitoring service, which gathers data from the local monitoring agents of the multiple devices (including sensors), process their information and forwards it to the optimisation modules; and modules for functional re-composition, which stitches together SW and HW functions and abstracts the changing of the operation of multiple devices to the performance optimisers, hence providing a technology-agnostic re-configurability service to the intelligent programs.

As an example of field-use of the proposed architecture, context-based system requirements could trigger a network-wide partitioning and reconfiguration of HW-accelerated and SW baseband functions, tailored for different traffic service delivery and QoE needs. Out of the different available options, the flexible partitioning could result in placing i) the eNB layer-2 and above together with the EPC functions in the Cloud (e.g., S1 traffic), ii) the eNB layer-2 and above functions at a local MEC server or distributed unit (DU) or iii) move the entire eNB stack in the Cloud (CPRI traffic) and transform the eNB in a Remote Radio Head (RRH). This dynamic split and reconfiguration of baseband functions is complementary to the partitioning of the eNB protocol stack, either at stack or algorithm level, e.g., MAC-PHY, Radio Link Control (RLC), Packet Data Convergence Protocol (PDCP) that was promoted recently by key industry actors in [50]. See also details on related RAN split options as considered by 3GPP in Section 6.6, and related deployment options in Section 6.7.

16.5.2 SW Platform Solutions for Prototyping 5G Systems

Some of the most relevant existing SW platforms that can support a vision such as the one described before are reviewed next. Several efforts have been done recently related to prototyping mobile networks in SW, with most of the platforms exploiting and supporting the GNU Radio development

suite [51] and the Ettus Research USRP SDR platforms [52][53]. One prominent product is Amari LTE 100, a fully SW-based LTE base station commercialized by Amarisoft [54] that offers a complete out-of-the-box solution for students and researchers. While being a relevant and promising product, with an outstanding computational efficiency, its closed license makes it unsuitable for MAC scheduling algorithms optimization and other advanced research fields.

The most popular open-source LTE SDR SW available for testbeds today are Eurecom's OpenAirInterface (OAI) [35] and openLTE [55]. OAI provides a standard-compliant implementation of a subset of Release 10 LTE, including key elements of the network such as UE, eNB, mobility management entity (MME), home subscriber server (HSS), serving gateway (S-GW) and packet gateway (P-GW) on standard Linux-based computing equipment (Intel x86 PC architectures). The SW can be used in conjunction with standard RF laboratory equipment available in many labs (i.e. National Instruments/Ettus USRP and PXIe platforms). Although OAI's implementation is very complete and provides relatively good performance, the code structure is complex and difficult to customize, plus the implementation of the CN functionality does not provide a very stable performance (although the pace of updates has increased recently).

openLTE is an open-source implementation of LTE specifications. The project includes a C library, Octave code for testing DL and UL physical random access channel (PRACH) functionalities, GNU Radio applications for DL functionalities, both simulated and using HW platforms, and a simple implementation of an eNB using Universal Software Radio Peripheral (USRP). It runs with the Ettus Research B2x0 USRP and provides eNB, MME and HSS functionalities. openLTE code is well organized, documented and easy to customize or modify. However, it is incomplete and many features are still unstable or under development. Furthermore, it does not provide an UE, limiting the testbed capabilities in terms of instrumentation and measurement.

A relatively recent library is srsLTE [37], an open-source platform for LTE experimentation designed for maximum modularity and code reuse and fully compliant with LTE Release 8, which can be considered as the evolution of the libLTE [56], an initial attempt that suffered from lack of functionality and poorly documented code. Although the platform provides a complete UE, this is not the case for the implementation of the eNB and Core Network, which are available for purchase, while the open version of the eNB lacks functionality such as hybrid automatic repeat request (HARQ).

Following the above overview, the OAI project by Eurecom [35] appears to be the most complete and flexible LTE SDR platform to date [57][58] and an implementation example will be described in the following sections targeting a vRAN/CRAN architecture.

16.6 Implementation Example: vRAN/C-RAN Architecture in OAI

As the wide adoption of the cloud computing concept consolidates, the currently distributed radio access network (D-RAN) architecture is expected to evolve toward a cloud/centralized radio access network (C-RAN) architecture that stands out as a promising solution for 5G. In C-RAN architecture, the original BS is decoupled into centralized baseband units (BBU) and the remote radio unit (RRU) at the network edge. These centralized BBUs can be pooled and used as shared resources, offering statistical multiplexing gains and energy efficiency. Further, C-RAN facilitates advanced coordinated multi-point (CoMP) processing and satisfies the stringent synchronization constraints of CoMP [59]. Finally, the BBU/RRU network functions can be implemented on commodity HW and

executed on a virtualized environment, i.e. virtualized RAN (vRAN), further benefiting from network softwarization and network function virtualization (NFV) concepts.

Despite its appeal, one key obstacle in the adoption of C-RAN is the excessive capacity requirements on the fronthaul (FH) link that provides BBU and RRU interconnections [60], as quantified for an example scenario in Section 6.6.2. To relax the excessive FH requirement, the concept of C-RAN has been revisited, and a more flexible distribution of baseband functionalities between the RRU and BBU is considered in [61] and also discussed in 3GPP. Rather than offloading all baseband processing to the BBU, it is possible to keep a subset of this processing at the RRU, which is the case for the 3GPP split options 7-x detailed in Section 6.2.2. This concept is also known as *flexible centralization* that splits the functions between RRU and BBU. Nevertheless, flexible centralization has two main drawbacks related to the initially envisioned benefits of C-RAN: (a) complex and expensive RRUs, and (b) reduction of the opportunities for multiplexing gains and coordinated processing. In consequence, flexible centralization is a trade-off between what is gained in terms FH requirements and what is lost in terms of C-RAN features.

As a consequence, a three-tier architecture is envisioned with the RRU at network edge, DU at some aggregation points and a central unit (CU) [62], see also Section 6.7. The low-level functional split between RRU and DU can maintain the low-cost RRUs, whereas the high-level functional split between DU and CU still provides multiplexing gain and supports advanced coordination schemes.

16.6.1 Overall Architecture

To support the envisioned three-tier architecture, OAI is evolving from an isolated monolithic BS concept (i.e., LTE eNB) to a more flexible and disaggregated BS spanning around multiple modules. The overall architecture supports multi-RAT, namely (e)LTE, NR and NB-IoT as shown in Figure 16-8. The envisioned RRU supports low-level L1 functionalities, while more signal processing is concentrated at the DU, when using low-latency FH link. Otherwise, more functions have to be distributed to RRUs, for example all the PHY layer may be placed at the RRU if 3GPP split option 6 is applied. Further, the CU covers a larger area than the DU and possesses a centralized PDCP functionality to allocate data-plane among different RATs within its coverage area. The same CU also hosts the control functions of the core network. All these three-tier components can be further managed by the global orchestrator under different deployment topologies. Please also refer to Section 6.7 for further details.

16.6.2 Deployment Topology

Since the C-RAN concept first appeared, a single FH link per RRU that connects all the way directly to the BBU has normally been assumed. However, it is expected that the FH network will evolve to a more complex multi-hop mesh network topology that requires switching and aggregation [63], see also Chapter 7. Hence, we focus on a multi-segment FH network topology in Figure 16-9. The considered topology can support a generic mesh deployment of the FH network that can be shared with other RRU traffic flows. The RRU gateways are the multiplex/de-multiplex point in the network and can be utilized to transport not only the C-RAN traffic, but also other traffic flows. Further, the BBU can be pooled and distributed centrally in the cloud, or at the network edge in some aggregated points. Figure 16-9 also shows how the considered C-RAN network can be mapped to the three-tier architecture stated before (i.e., RRU, DU, CU) and the ones provided by Next Generation Fronthaul Interface (NGFI), such as the RRU, the radio aggregation unit (RAU) and radio cloud center (RCC) in [64], or

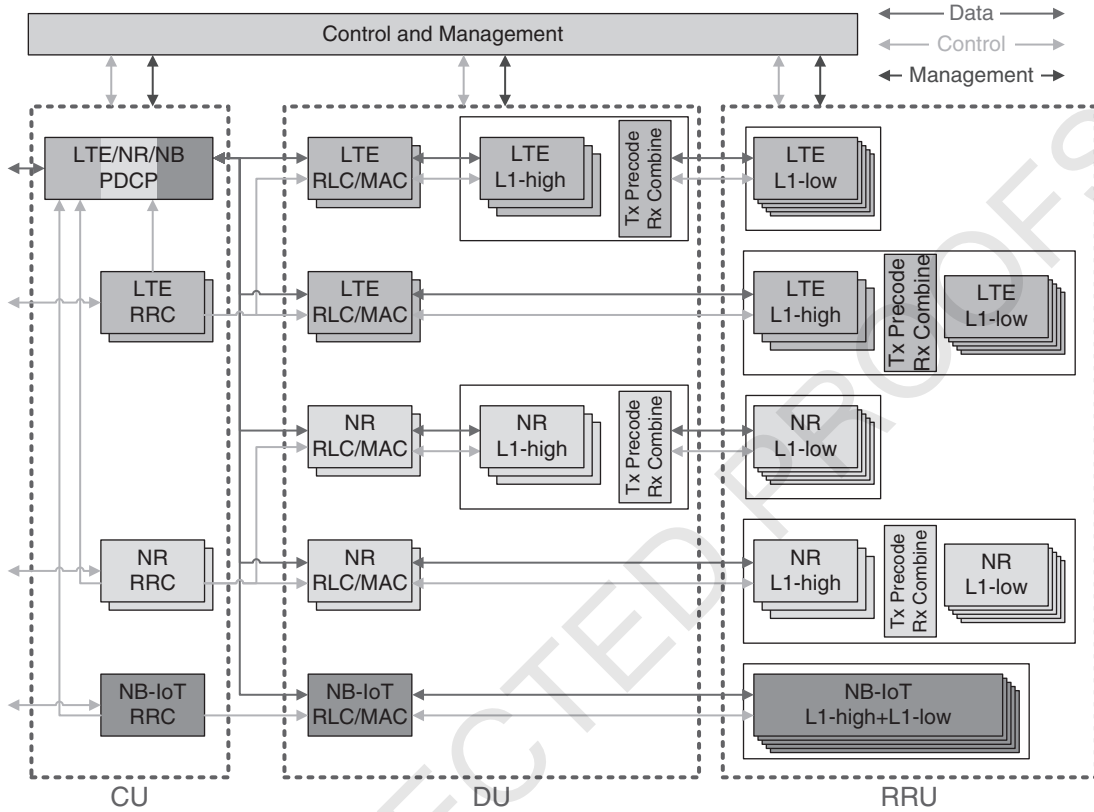


Figure 16-8. OpenAirInterface (OAI) three-tier heterogeneous RAN architecture, see also Section 6.7.

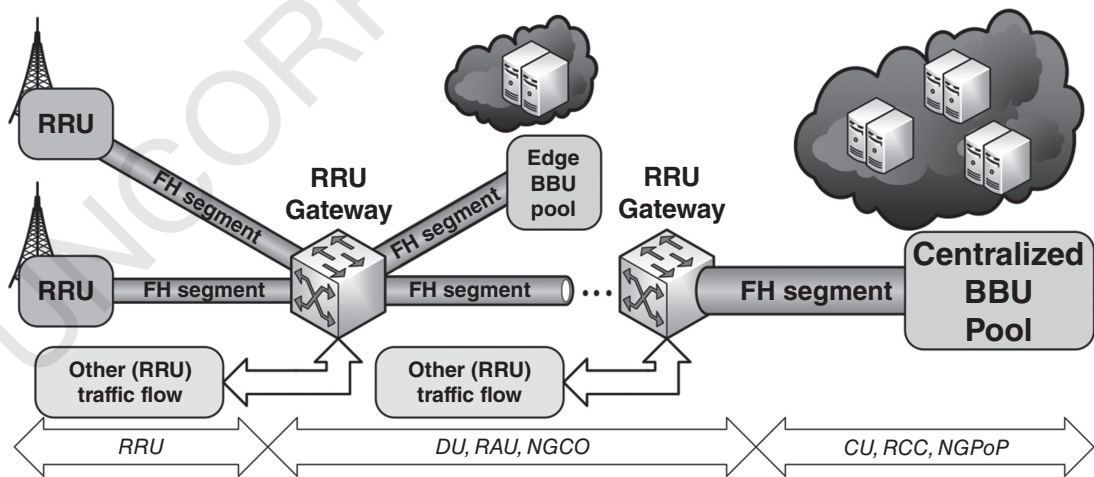


Figure 16-9. Considered C-RAN network topology.

the telecom industry, which considered an RRU, new generation central office (NGCO), and a next generation point of presence (NGPoP). Another key characteristic is how the information between the RRU and BBU is transported over the FH link. A number of FH transmission protocols are already used in the field, such as CPRI [65], OBSAI [66] and ORI [67]. However, as stressed before, these techniques consider carrying raw I/Q samples that can be utilized in the FH segment between first and second tier of the architecture in Figure 16-9. In light of the flexible centralization concept, different types of information are transported over the FH link based on the functional split between RRU and BBU, such as the 3GPP split options 7-x that are described in detail in Section 6.6.2. Given the extensive adoption of Ethernet in remote clouds, data centers and the core network, the Radio over Ethernet (RoE) [68] approach is a generic, cost-effective, off-the-shelf alternative for FH link traffic transport.

16.6.3 Performance Results

We evaluate the C-RAN implementation using OAI [35], a SW-based LTE/LTE-A system implementation spanning the full 3GPP protocol stack with a third-party EPC, commercial off-the-shelf (COTS) UE and USRP B210 SDR. In the following, we consider two C-RAN network deployments: (a) no RRU gateway (i.e., 1 hop between RRU/DU) and (b) single RRU gateway (i.e., 2 hops between RRU/DU) with each FH link being made up of a 3-meter cable. In simple, only 1 RRU and 1 DU in the considered C-RAN network using split A or split B defined in [69] with 5 MHz/10 MHz radio bandwidth. In the following, some important KPIs related to C-RAN are presented.

16.6.3.1 FH-related KPIs: FH link Throughput

The FH link throughput together with the theoretical rate of both 5 MHz and 10 MHz cases are shown in Figure 16-10. First, the RAW transmission throughput only has little overhead (between 3 to 4 Mbps) compared with the theoretical rates. The UDP transportation shows little further overhead (up to 3 Mbps) compared with RAW transmission. Moreover, the applied a-law compression scheme reaches almost 50 % reduction in the FH link throughput. Further, using split B shows the gain of 43.8 % in terms of FH link throughput reduction compared with split A via moving the FFT/IFFT operation to RRU, as considered in 3GPP split option 7-1. This reduction ratio is similar to that presented in Section 6.2.2, and also close to the analysis in [60] that shows 45.3 % of throughput reduction. We can also observe that the throughput is scaling with the channel bandwidth and hence the predicted FH capacity when using 20 MHz channel bandwidth without any compression scheme will be around 1 Gbps.

16.6.3.2 FH-related KPIs: Round-trip Time of the FH and RF Front-end

These KPIs measure the round-trip latency between the FH link and RF devices. Such metrics are crucial to evaluate the possible RRU-DU functional split, and for instance NGMN adopts 250 μ s as the maximum one-way FH latency [70] and SCF categorizes the one-way FH latency from 250 μ s to the millisecond level to evaluate the applicable RRU-DU split [71]. Here, we use the difference of timestamps at RRU side to measure both round-trip times (RTTs). The RTT of the FH is measured as the time elapsed at the RRU between the start of sending the receiver data samples and the end of reading the corresponding transmitter data samples from the DU on the FH link. In detail, this RTT of the FH is made up of 5 components: (a) compress time, (b) FH write time, (c) FH link RTT, (d) FH read time, and (e) decompress time. Moreover, the RTT of

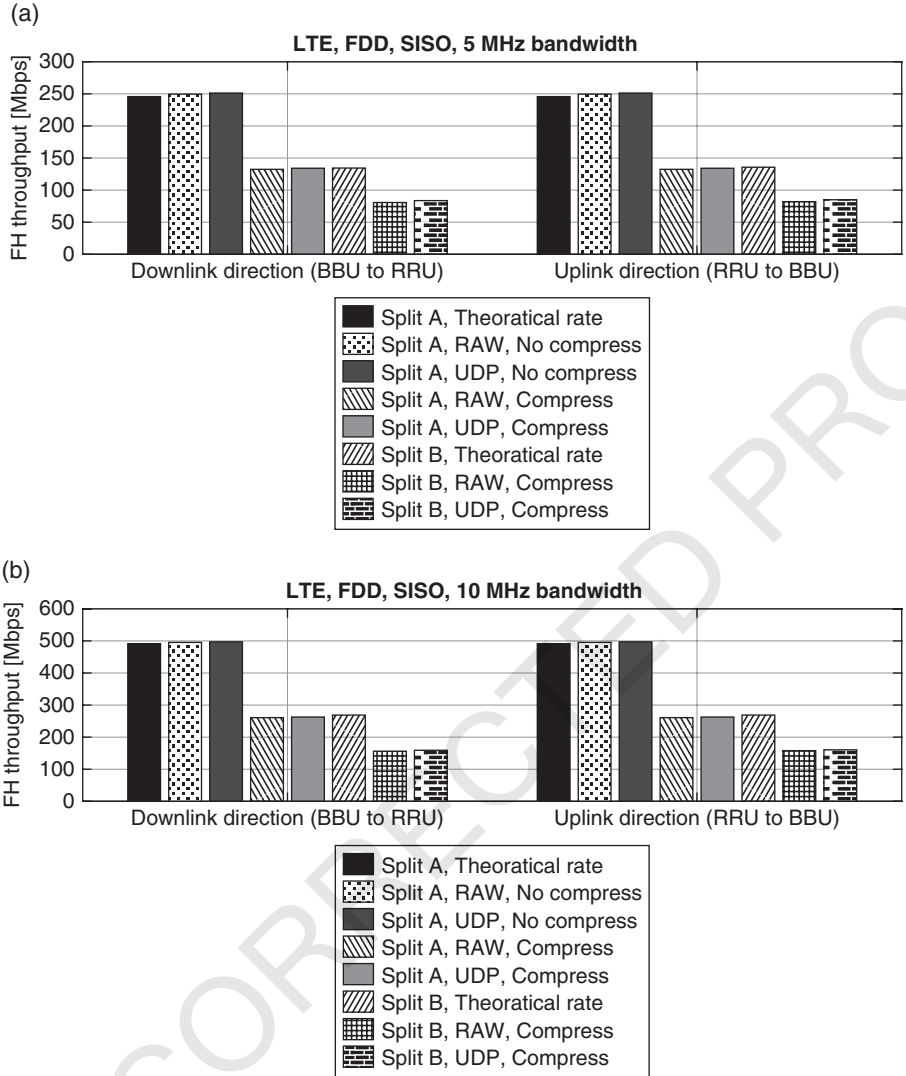


Figure 16-10. FH throughput needed for 5 MHz and 10 MHz bandwidth.

the RF front-end is defined as the time elapsed from the reading of data samples of RF devices until the writing of samples to RF devices, which includes the RTT of the FH.

Based on aforementioned definitions, the results are shown in Figure 16-11. The reduction of the RTT in the FH is proportional to the reduction of throughput (by a factor of 2) when applying the a-law compression which comes at the cost of the extra processing time for compression and decompression (see Table 16-2) but with much fewer FH link read/write time (see Table 16-2), which confirms the benefit of the compression in the FH network. In addition, it can be seen from Figure 16-11(b) that the average RTT of the RF front-end remains comparable for the 5 MHz case even without the

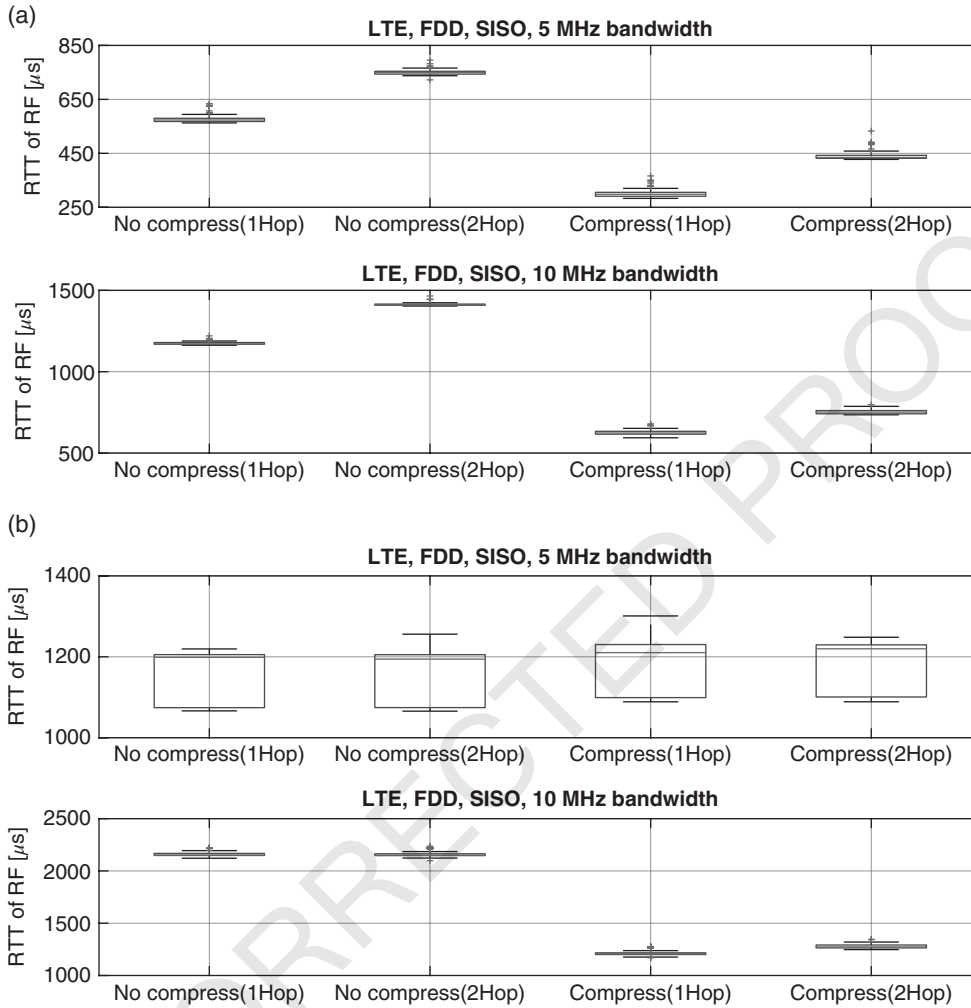


Figure 16-11. RTT of the FH and RF for 5 MHz/10 MHz bandwidth.

Table 16-2. Mean time for FH link read/write and compression/decompression.

Bandwidth	Compression	Compress time [μs]	FH write time [μs]	FH read time [μs]	Decompress time [μs]
5 MHz	No	-	69.19	224.37	-
	Yes	16.53	68.80	53.13	23.43
10 MHz	No	-	72.35	469.53	-
	Yes	31.70	71.95	170.13	35.93

compression since the RTT of FH is less than the duration of one transmission time interval (TTI, 1 ms in this case) in Figure 16-11(a). However, in case of 10 MHz, the RTT of the FH in Figure 16-11(a) without compression is significantly larger than 1 ms, corresponding to the duration of one TTI, which in turn greatly increases the RTT of RF front-end in Figure 16-11(b).

16.6.3.3 Endpoint-related KPIs

The end-point HW load at the RRU/DU comprises the central processing unit (CPU) and memory utilization. Such measurement can be utilized for two purposes: (a) estimate the number of RRUs that can be supported under the limited number of DUs in the pool given the fixed functional split or (b) dynamic functional split based on the HW load to fully utilize all available resources among RRU and DU. Hence, we compare D-RAN and C-RAN deployments given the constant user traffic of 15 Mbps/30 Mbps in downlink and 5 Mbps/10 Mbps in uplink for 5 MHz/10 MHz bandwidth. The results are listed in Table 16-3, where the CPU utilization ratio is the percentage of CPU processing time of the process and the memory usage is measured based on the proportional set size (PSS) in KBytes. RRU, DU and eNB are deployed in a hex-core CPU each with Intel i7 Sandy Bridge architecture in 3.2 GHz. As a result, 2 CPU cores are required to deploy the proposed RRU for 10 MHz bandwidth for split A and 3 cores for split B. Moreover, the sum of CPU resources required by RRU and DU is slightly higher than the one required by the eNB deployment. In addition, the memory usage at RRU and DU does not have large differences since our C-RAN deployment is considered to support the fully flexible function split, i.e., the baseband processing can be dynamically allocated between RRU and DU. Hence, all baseband functionalities are still at both RRU and DU. We also observe that if we only deploy necessary functionalities based on the functional split, i.e. split-specific deployment, the CPU ratio and memory usage are largely reduced, for instance, to occupy 5 % of CPU ratio and 16 KBytes of memory usage for RRU in split A.

16.6.3.4 User-plane KPIs

We first clarify the relations between user plane (UP) and FH-related KPIs. Taking the FH delay as an example, it can be absorbed and compensated by scheduling the transmission ahead of time, which in

Table 16-3. HW load of RRU/DU.

Bandwidth	Split	Endpoint	CPU ratio	Memory usage [kBytes]
5 MHz	eNB		40.15 %	1002019
	A	RRU	16.76 %	917486
		DU	26.57 %	918794
	B	RRU	24.19 %	917478
		DU	22.71 %	917174
10 MHz	eNB		65.02 %	1195059
	A	RRU	29.23 %	1107382
		DU	45.70 %	1180126
	B	RRU	41.12 %	1107374
		DU	32.40 %	1124989

turn reduces the total transmitter and receiver processing time to provide extra FH transportation time. However, such shortened processing time might not be enough for some processing (e.g., turbo decoder) in some cases [72], and it can cause the extra UP delay due to retransmission. To evaluate the UP KPIs, we use 15 Mbps/30 Mbps user traffic for 5 MHz/10 MHz in downlink direction separately.

To characterize the UP QoS from the user perspective, we measure the delay jitter and good-put. First, the delay jitter is shown in Figure 16-12. Due to the extra route from the user to the gateway that leads to less available transmitter/receiver processing time, the jitter of the C-RAN deployment is larger than the one in a legacy D-RAN, and the jitter of the case with two hops is larger than the one in one hop case. Further, the measured good-put at application level is shown in Figure 16-13, and seven different C-RAN deployments (A1 to A5, B1, and B2 in Table 16-4) are considered to be compared to a legacy eNB/D-RAN deployment. Note that the good-put is the application-level successful throughput that is significant from the user perspective. We observe that these deployment scenarios show almost the same good-put variation as the D-RAN one; that is to say, the experienced good-put will be maintained among different RAN deployments.

For the UP delay, we measure the application RTT over the default radio bearer. It not only considers the impact of the FH link, but also the transmitter and receiver processing time at the RRU and DU of both downlink and uplink directions. Hence, we use the ping utility with a 8192 bytes packet size and 0.2s inter-departure time to characterize the delay in Figure 16-14. We observe that the average RTT in a

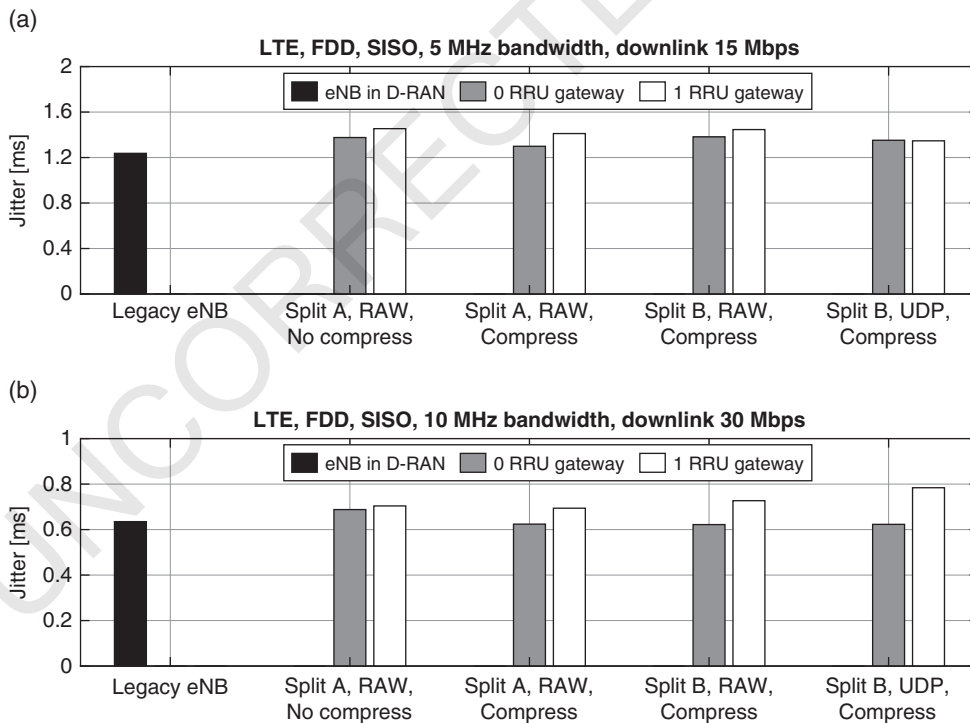


Figure 16-12. Packet delay jitter for 5 MHz/10 MHz bandwidth.

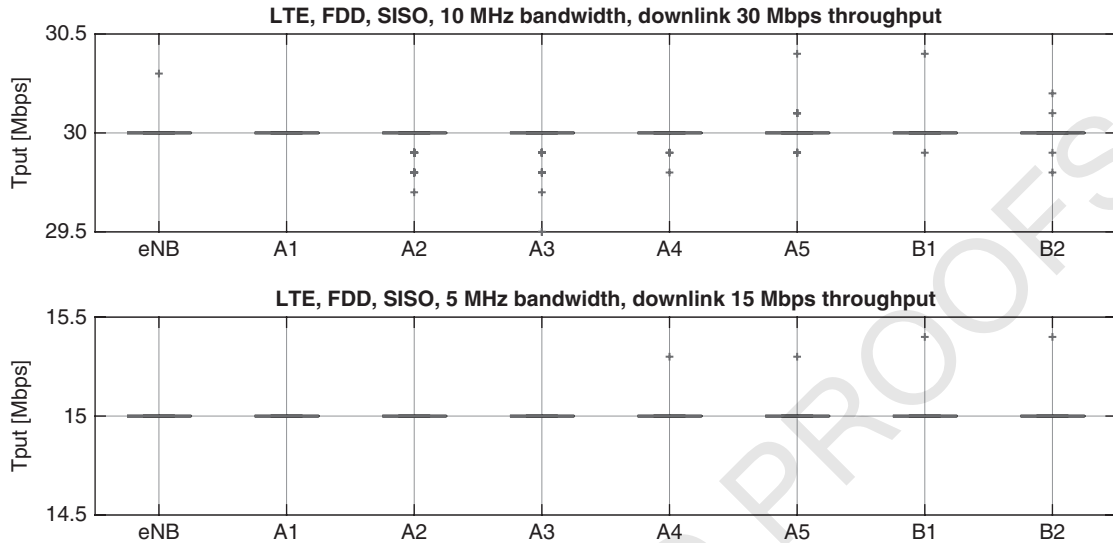


Figure 16-13. Data-plane good-put of several deployment scenarios.

Table 16-4. Parameters for the different C-RAN deployments.

Mode	Split	Protocol	Compression	Hop count
A1	A	RAW	No	1
A2	A	RAW	No	2
A3	A	UDP	No	2
A4	A	RAW	Yes	2
A5	A	UDP	Yes	2
B1	B	RAW	Yes	2
B2	B	UDP	Yes	2

C-RAN deployment is a little higher than the one in a legacy eNB/D-RAN deployment. However, in the two-hop cases (i.e., A2 to A5, B1, B2), the long tail distribution is exhibited due to the extra route from the user to the gateway that reduces the time or transmitter/receiver processing. In this sense, once the processing cannot be finished within the available time then the re-transmission scheme will increase the data-plane delay.

16.6.4 Deployment Environment

An RRU prototype comprises all necessary components is shown in Figure 16-15. It contains the Pico-ITX or other smaller motherboard (e.g., UpBoard with Intel Atom quad-core processor), Power over Ethernet (PoE+) to support power supply, wiring for 1 Gigabit/sec Ethernet, RF front-end components (PA, LNA, Switch), 10 MHz/PPS frequency synchronization cable, baseband-to-RF radio unit (e.g., USRP B200-mini) and RF front-end circuits.

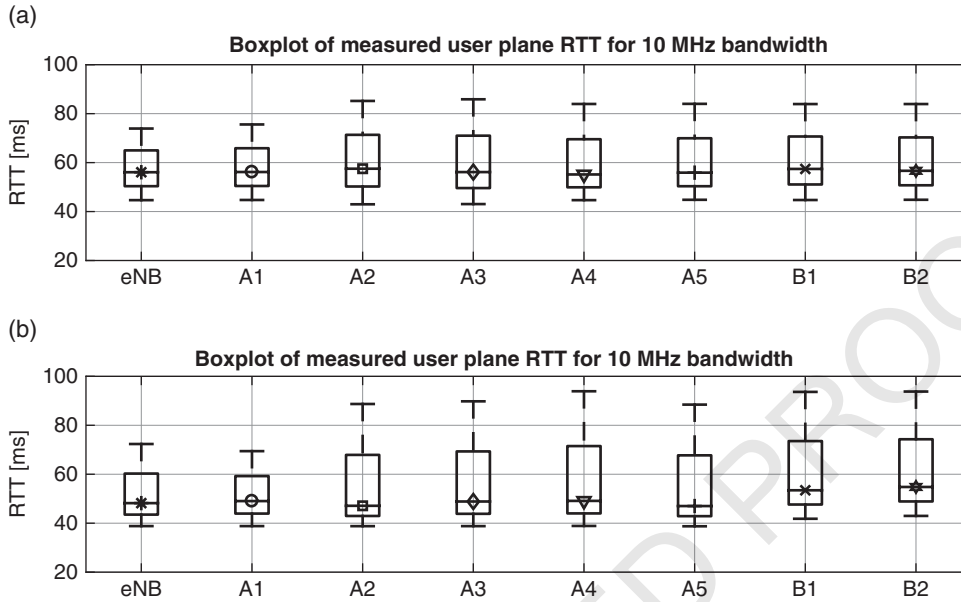


Figure 16-14. User plane packet RTT for 5 MHz and 10 MHz BW.

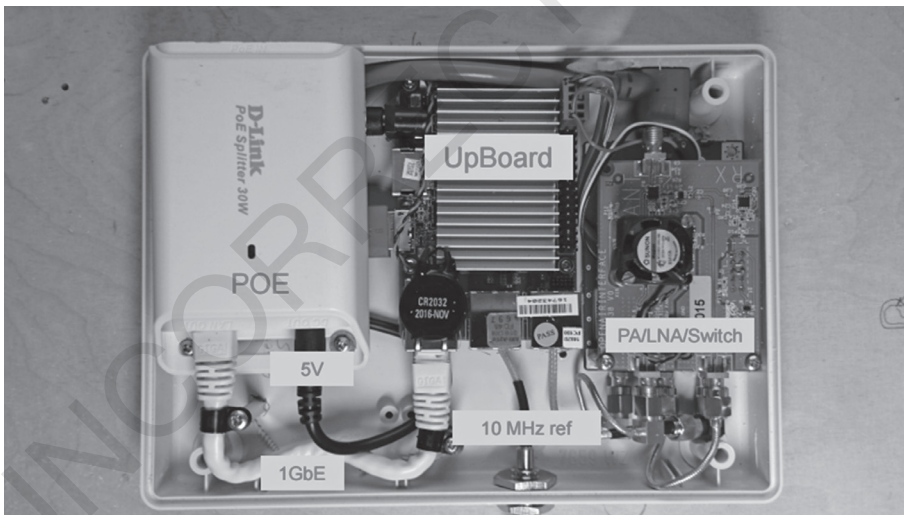


Figure 16-15. RRU prototype.

In Figure 16-16, the physical deployment of an indoor scenario is shown. Two distribution switches are placed and connected to the aggregation switch. As an extension, the planned outdoor deployment using a large number of RRUs (i.e. up to 64 RRU elements) and higher FH capacity is depicted in Figure 16-17. This deployment showcases the coordination of the indoor and outdoor network segments and the orchestration on a larger-scale.

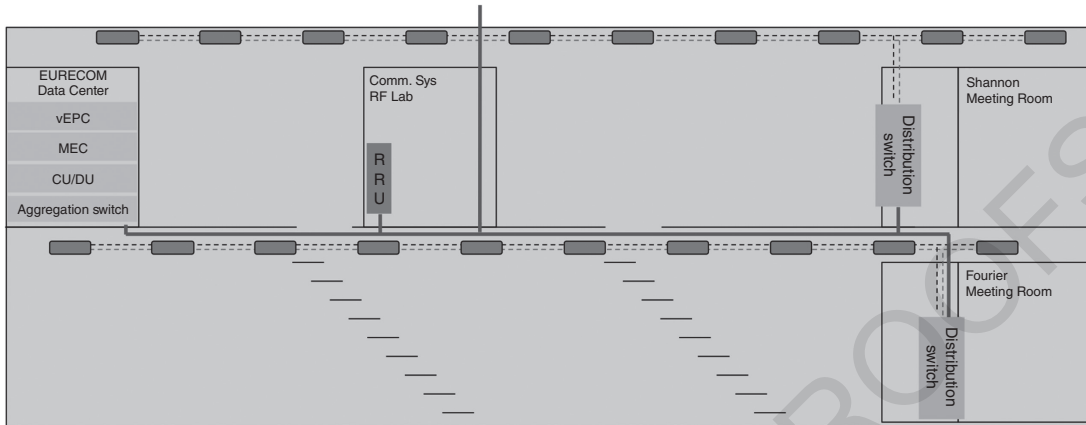


Figure 16-16. Indoor deployment floorplan.

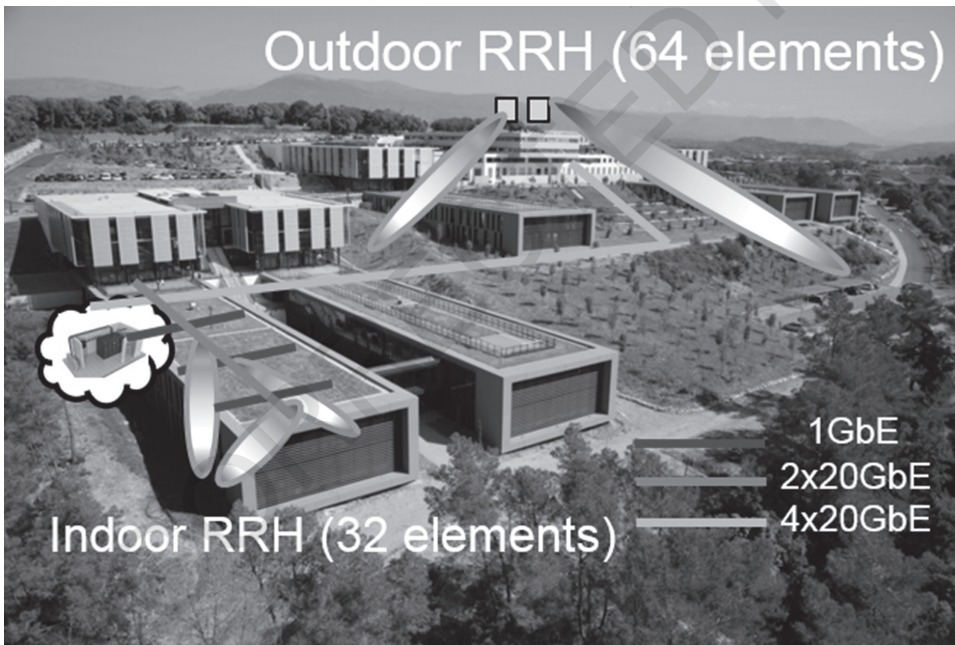


Figure 16-17. Campus outdoor deployment.

16.7 Summary

In this chapter, a number of implementation challenges that future 5G HW and SW platforms will face have been presented, and some preliminary solutions in different technology domains have been provided. In particular, Section 16.2 has provided an overview on challenges and solutions regarding

the analog HW and mixed signal stages, whereas in Section 16.3 digital HW has been addressed with a strong emphasis on waveform implementation complexity. Section 16.4 has focused on HW/SW function splitting and a set of functional requirements for SW platforms has been described in Section 16.5 together with a summary of different existing SW platforms which could support these. As an example of one of such platforms, in Section 16.6, an implementation targeting a vRAN/CRAN architecture has been provided.

References

- 1 ETSI TS 136 104, “LTE; Evolved Universal Terrestrial Radio Access E-UTRA); Base Station (BS) radio transmission and reception (3GPP TS 36.104 version 12. Release 12)”, V12.5.0, Oct. 2014
- 2 5G PPP Flex5Gware project, Deliverable D2.2, “Analogue components for high-performing and versatile 5G RF front-ends”, May 2017
- 3 5G PPP Flex5Gware project, Deliverable D3.2, “Mixed-signal strategies for 5G: final architecture and design for flexibility, power and spectral efficiency”, May 2017
- 4 Analog Devices, AD9172, see <http://www.analog.com/media/en/technical-documentation/data-sheets/AD9172.pdf>
- 5 Vadatech, AMC589, see http://www.vadatech.com/media/AMC589_AMC589_Datasheet.pdf
- 6 W.H. Doherty, “A new high efficiency power amplifier for modulated waves”, Proceedings of the Institute of Radio Engineers, vol. 24, no. 9, pp. 1163–1182, Sept. 1936
- 7 M. Bozzi, A. Georgiadis and K. Wu, “Review of substrate-integrated waveguide circuits and antennas”, in IET Microwaves, Ant. & Prop., vol. 5, no. 8, pp. 909–920, June 2011
- 8 S. Agneessens and H. Rogier, “Compact Half Diamond Dual-Band Textile HMSIW On-Body Antenna”, in IEEE Trans. on Antennas and Propagation, vol. 62, no. 5, May 2014
- 9 M. Daniel, Y. Xin, H. Heimpel and G. Fischer, “An All-Digital, Single- Bit RF Transmitter for Massive MIMO”, IEEE Transactions on Circuits and Systems, vol. 64, no. 3, pp. 696–704, Dec. 2016
- 10 5G PPP Flex5Gware project, Deliverable D6.2, “Final PoC evaluation in Flex5Gware”, June 2017
- 11 Goyal, S. et al., “Improving Small Cell Capacity with Common-Carrier Full Duplex Radios”, IEEE International Conference on Communications (ICC 2014), June 2014
- 12 E. Costa, M. Midro and S. Pupolin, “Impact of amplifier nonlinearities on OFDM transmission system performance”, IEEE Comm. Let., vol. 3, pp.37–39, Feb. 1999
- 13 P. Achaichia, M. L. Bot and P. Siohan, “Windowed OFDM versus OFDM/OQAM: A transmission capacity comparison in the HomePlug AV context”, IEEE Intl. Symp. on Power Line Communications and Its Applications (ISPLC 2011), Apr. 2011
- 14 G. Berardinelli, F. M. L. Tavares, T. B. Sørensen, P. Mogensen and K. Pajukoski, “On the potential of zero-tail DFT-spread-OFDM in 5G networks”, IEEE Vehicular Technology Conference (VTC Fall 2014), Sept. 2014
- 15 P. Siohan, C. Siclet and N. Lacaille, “Analysis and design of OFDM/OQAM systems based on filterbank theory”, IEEE Transactions on Signal Processing, vol. 50, no. 5, pages 1170–1183, 2002
- 16 C. Kim, Y. H. Yun, K. Kim and J. Y. Seol, “Introduction to QAM-FBMC: From waveform optimization to system design”, IEEE Comm. Magazine, vol. 54, no. 11, pages 66–73, 2016
- 17 Z. Zhao, M. Schellmann, Q. Wang, X. Gong, R. Boehnke and W. Xu, “Pulse shaped OFDM for asynchronous uplink access”, 49th Asilomar Conference on Signals, Systems and Computers, Nov. 2015

- 18 L.G. Baltar, F. Schaich, M. Renfors and J.A. Nossek, "Computational complexity analysis of advanced physical layers based on multicarrier modulation", in Future Network & Mobile Summit (FutureNetw 2011), June 2011
- 19 M. Bellanger et al., "FBMC physical layer: a primer", PHYDYAS, Jan. 2010
- 20 M. Bellanger, "Specification and design of a prototype filter for filter bank based multicarrier transmission", IEEE International Conference on Acoustics, Speech, and Signal Processing, May 2001
- 21 T. Wild and F. Schaich, "A Reduced Complexity Transmitter for UF-OFDM", Vehicular Technology Conference (VTC Spring 2015), May 2015
- 22 Y. Chen, F. Schaich and T. Wild, "Multiple Access and Waveforms for 5G: IDMA and Universal Filtered MultiCarrier", Vehicular Technology Conference (VTC Spring 2014), May 2014
- 23 N. Michailow, I. Gaspar, S. Krone, M. Lentmaier and G. Fettweis, "Generalized frequency division multiplexing: Analysis of an alternative multi-carrier technique for next generation cellular systems", International Symposium on Wireless Communication Systems (ISWCS), Aug. 2012
- 24 I. Gaspar, N. Michailow, A. Navarro, E. Ohlmer, S. Krone and G. Fettweis, "Low Complexity GFDM Receiver Based on Sparse Frequency Domain Processing", Vehicular Technology Conference (VTC Spring 2013), June 2013
- 25 5G PPP Flex5Gware project, Deliverable D4.2, "Final report on HW architectures", May 2017
- 26 3GPP TS 38.802, "Study on New Radio Access Technology - Physical Layer Aspects", March 2017
- 27 E. O. Brigham, "The fast Fourier transform", Prentice-Hall Inc., 1998
- 28 5G PPP METIS-II project, Deliverable D4.2, "Final air interface harmonization and user plane design", April 2017
- 29 M. Forsell and J. Roivainen, "REPLICA T7-16-128 - A 2048-threaded 16-core 7-FU chained VLIW chip multiprocessor", in Special session on Multicore, Manycore and Distributed systems at the 48th Asilomar Conference on Signals, Systems, and Computers, Nov. 2014
- 30 M. Forsell, J. Roivainen and V. Leppänen, "Prototyping the MBTAC processor for the REPLICA CMP", 16th Workshop on Advances in Parallel and Distributed Computational Models (APDCM'14) in conjunction with the 28th IEEE International Parallel and Distributed Processing Symposium (IPDPS'14), May 2014
- 31 T. Richardson, A. Shokrollahi and R. Urbanke, "Design of capacity-approaching irregular low-density parity-check codes", IEEE Transactions on Information Theory, vol. 47, no. 2, pp. 619–637, 2001
- 32 R. Tanner, "A recursive approach to low complexity codes", IEEE Transactions on Information Theory, vol. 27, no. 5, pp. 533–547, 1981
- 33 Truong Nguyen-Ly, Khoa Le, V. Savin, D. Declercq, F. Ghaffariy, and O. Boncalo, "Non-Surjective Finite Alphabet Iterative Decoders", IEEE International Conference on Communications (ICC 2016), May 2016
- 34 LTE-EPC Network Simulator (LENA), see [http://iptechwiki.cttc.es/LTE-EPC Network Simulator \(LENA\)](http://iptechwiki.cttc.es/LTE-EPC%20Network%20Simulator%20(LENA))
- 35 N. Nikaein, R. Knopp, F. Kaltenberger, L. Gauthier, C. Bonnet, D. Nussbaum, and R. Ghaddab, "OpenAirInterface 4G: an open LTE network in a PC", 2013
- 36 Q. Zheng, H. Du, J. Li, W. Zhang and Q. Li, "Open-LTE: An Open LTE simulator for mobile video streaming", IEEE International Conference on Multimedia and Expo Workshops (ICMEW 2014), July 2014
- 37 I. Gomez-Migueluez, A. Garcia-Saavedra, P. D. Sutton, P. Serrano, C. Cano and D. J. Leith, "srsLTE: An Open-Source Platform for LTE Evolution and Experimentation", Feb. 2016
- 38 5G PPP Flex5Gware project, Deliverable D4.2, "Final report on HW architectures", May 2017

- 39 C. J. Bernardos et al., “An Architecture for Software Defined Wireless Networking”, *IEEE Wireless Communications*, vol. 21, no. 6, pp. 52–61, June 2014
- 40 C. Donato et al., “An OpenFlow Architecture for Energy Aware Traffic Engineering in Mobile Networks”, *IEEE Network*, vol.29, no.4, pp. 54–60, July–August 2015
- 41 P. Demestichas, G. Dimitrakopoulos, J. Strassner and D. Bourse, “Introducing Reconfigurability and Cognitive Networks Concepts in the Wireless World”, *IEEE Vehicular Technology Magazine*, 2006
- 42 S. Hong, J. Mehlman and S. Katti, “Picasso: Flexible RF and Spectrum Slicing”, *ACM SIGCOMM*, Aug. 2012
- 43 M. Mueck, et al., “ETSI Reconfigurable Radio Systems: Status and Future Directions on Software Defined Radio and Cognitive Radio Standards”, *IEEE Communications Magazine*, vol. 48, no. 9, Sept. 2010
- 44 C. R. Aguayo González, C. B. Dietrich and J. H. Reed, “Understanding the Software Communications Architecture”, *IEEE Communications Magazine*, vol. 47, no. 9, Sep. 2009
- 45 J. Bard and V. J. Kovarik Jr., “Software Defined Radio: The Software Communications Architecture”, John Wiley & Sons Ltd., 2007
- 46 P. De Mil, B. Jooris, L. Tytgat, J. Hoebeke, I. Moerman and P. Demeester, “SnapMac: A generic MAC/PHY architecture enabling flexible MAC design”, *Ad Hoc Networks*, 2014
- 47 I. Tinnirello, G. Bianchi, P. Gallo, D. Garlisi, F. Giuliano and F. Gringoli, “Wireless MAC Processors: Programming MAC Protocols on Commodity Hardware”, *IEEE INFOCOM*, Mar. 2012
- 48 G. Bianchi, P. Gallo, D. Garlisi, F. Giuliano, F. Gringoli and I. Tinnirello, “MAClets: Active MAC Protocols over Hard-Coded Devices”, *ACM CONEXT*, Dec. 2012
- 49 P. D. Sutton et al., “Iris: An Architecture for Cognitive Radio Networking Testbeds”, *IEEE Communications Magazine*, vol. 48, no. 9, Sept. 2010
- 50 Small Cell Forum, “Virtualization for small cells: Overview”, see http://www.scf.io/en/documents/106_Virtualization_for_small_cells_Overview.php
- 51 GNU Radio, “GNU Radio: The Free & Open Software Radio Ecosystem”, see <http://gnuradio.org>
- 52 Ettus Research, see <http://www.ettus.com/>
- 53 M. Abirami, V. Hariharan, M. Sruthi, R. Gandhiraj and K. Soman, “Exploiting GNU Radio and USRP: An Economical Test Bed for Real Time Communication Systems”, *IEEE ICCCN*, July 2013
- 54 Amarisoft, “Amari LTE 100 - Software LTE base station on PC”, see <http://www.amarisoft.com/index.php?p=amarilte>
- 55 B. Wojtowicz, “openLTE: An open source 3GPP LTE implementation”, last update Sept. 2014, see <http://sourceforge.net/projects/openlte/>
- 56 I. Gomez, “libLTE: Open source 3GPP LTE library”, last update Oct 2013, see <http://sourceforge.net/projects/liblte/>
- 57 R. Wang, Y. Peng, H. Qu, W. Li, H. Zhao and B. Wu, “OpenAirInterface-An Effective Emulation Platform for LTE and LTE-Advanced”, *IEEE ICUFN*, 2014
- 58 H. Anouar, C. Bonnet, D. Cámara, F. Filali and R. Knopp, “An Overview of OpenAirInterface Wireless Network Emulation Methodology”, *ACM SIGMETRICS Perform. Eval. Rev.*, 2008
- 59 A. Checko et al., “Cloud RAN for mobile networks - a technology overview”, *IEEE Communications Surveys & Tutorials*, 2014
- 60 C.-Y. Chang et al., “Impact of packetization and functional split on C-RAN fronthaul performance”, *IEEE International Conference on Communications, (ICC 2016)*, May 2016
- 61 D. Wübben et al., “Benefits and impact of cloud computing on 5G signal processing: Flexible centralization through cloud-RAN”, *IEEE Signal Processing Magazine*, vol 31, no. 6, pp. 35–44, Oct. 2014

- 62 I. Chih-Lin, "RAN revolution with NGFI (xhaul) for 5G", Optical Fiber Communications Conference, March 2017
- 63 C.-Y. Chang et al., "Impact of packetization and scheduling on C-RAN fronthaul performance", IEEE Global Communications Conference (GLOBECOM 2016), Dec. 2016
- 64 Y. Zhiling et al., White Paper, "White paper of next generation fronthaul interface v1.0", China Mobile Research Institute, Tech. Rep., 2015
- 65 CPRI, "Interface specification v7.0", 2015
- 66 OBSAI, "BTS system reference document v2.0", 2006
- 67 ETSI GS ORI 001, "Open Radio equipment Interface (ORI); requirements for ORI", Oct. 2014
- 68 IEEE, Technical Report, "1904.3 task force: Standard for radio over Ethernet encapsulations and mapping", 2015
- 69 C.-Y. Chang et al., "FlexCRAN: A Flexible Functional Split Framework over Ethernet Fronthaul in Cloud-RAN", IEEE International Conference on Communications (ICC 2017), May 2017
- 70 NGMN, Technical Report, "Further study on critical C-RAN technologies", 2015
- 71 Small Cell Forum, Technical Report, "Small cell virtualization functional splits and use cases", 2015
- 72 N. Nikaein, "Processing radio access network functions in the Cloud: Critical issues and modeling", Intl. Workshop on Mobile Cloud Computing and Services, Sept. 2015