On the feasibility of MAC and PHY split in Cloud RAN

Ghizlane Mountaser, Maria Lema Rosas, Toktam Mahmoodi, Mischa Dohler
Centre for Telecommunications Research
Department of Informatics, King’s College London, UK
{ghizlane.mountaser, maria.rosas, toktam.mahmoodi, mischa.dohler}@kcl.ac.uk

Abstract—Splitting functionalities of radio access network (RAN) and cloudification of such functionalities is considered as one of the key enablers of the next generation mobile and wireless networking, i.e. 5G, and is often referred to as software-defined RAN, virtualized RAN or Cloud RAN. Defining the splitting point, and maintaining the tight interaction between different functionalities in the RAN is, however, critical. Success of such cloudification depends on the availability of high speed fronthaul, while high speed fronthauling is costly. In this paper we experiment splitting MAC and PHY layer with fronthauling through Ethernet that allows using commodity and low-cost industry standard equipment. We examine the effect of packetization on latency, and study the pros and cons of splitting MAC and PHY layer, within a hardware-based testbed.

Index terms— C-RAN;Cloud RAN; layer split; fronthaul; 5G; latency;

I. INTRODUCTION

The fifth generation of mobile communication networks (5G) is facing different multiple challenges compared to previous generations, with an ever increasing number of use cases, it is intended to satisfy all the new applications being considered by so-called industry verticals. Besides giving users mobile broadband services, 5G is also expected to provide technological solutions for time sensitive communications with the rising of different vertical domains, such as industrial networking [1], the Tactile Internet [2], or various scenarios within Internet of Things [3]. The wide range of services being provided is changing the paradigm of cellular networks, and concepts as common as cells are no longer relevant. In fact, 5G is evolving to a device-centric approach [4], [5], where the configurability (and reconfigurability) of the network is key to satisfy the quality of service (QoS) [6].

To support this new type of network, trends like softwareization and centralization are being widely considered by both the research community and standardisation bodies [7], [8]. In particular, centralization in the radio access network (RAN) has been discussed in the context of Cloud or Centralised RAN (C-RAN) [9], [10], where the base band processing unit (BBU) is decoupled from the remote radio head (RRH). Various advantages are enabled through centralization of RAN functions including enhanced cooperative solutions, improved load balancing and RAN sharing, among others. On the other hand, it can introduce huge challenges in delivering low-latency services, and requires high bandwidth availability to transport the base band signals. To reduce bandwidth requirements, some low layer network functions may remain close to the RRH, reducing the level of centralization. However, such split of network functions depend largely on the availability of transport networks, and the final RAN configuration (i.e., distributed or centralised) will determine the level of cooperation and the deployment of some of the features being considered in the road to 5G. Linked to this, work in [11] surveys all transport network solutions available for fronthaul and discusses the impact of such technologies in the RAN context. In particular, how centralised RAN network functions impacts the service requirements, but simultaneously impacts the level of cooperation among different transmitters.

To this end, the an interesting research questions arises to evaluate the feasibility of these different levels of centralization, also called functional split. This question has been addressed in the literature from both theoretical and implementation perspectives. For example, the work presented in [12] reviews the main challenges to support different levels of centralization, and discusses technical solutions like compression, quantisation and RRH clustering to support fully centralised RAN. The impact of different functional split in terms of overhead data in the fronthaul is discussed in [13]. Furthermore, the bandwidth requirements for supporting different radio configurations (for instance, carrier bandwidth and number of antennas) are studied [14], [9], where it has been shown that including more antennas can increase the required bandwidth to hundreds of Gbps.

On the other hand, there is a strong interest from telecom industry to leverage packet switched networks, such as Ethernet, in providing a cost-effective transport network solution for C-RAN. This will allow the use of lower cost-industry standard equipment and sharing infrastructure already deployed for fixed networks. The most widely used transport protocol between the central entity and the remote unit is the Common Public Radio Interface (CPRI), which has been specifically designed based on the requirements of digitised base band signals. However, Ethernet is a best effort based technology, and it is not designed to meet the low jitter and latency requirements for base band signals transmission, i.e., CPRI. Therefore, works considering CPRI over Ethernet [14], [15] suggest providing dedicated links between RRH and BBU, and enhancing the Ethernet network with additional features to satisfy stringent latency and jitter constraints. In this context, allowing for a higher layer split can allow the
use of packet switched networks without degrading the overall RAN performance. The effect of packetization in different C-RAN splits in terms of latency and overhead is studied in [16], where different functional splits are simulated. Number of packetization methods are also examined in order to support the tight RAN deadlines and the results focus on demonstrating effect of latency and overhead on the fronthaul performance.

This paper focuses on feasibility study of medium access control (MAC) and physical (PHY) layer split, over the Ethernet and steps forward by taking this study into implementation. We evaluate the impact of packetization in a MAC-PHY split using Software Defined Radio (SDR) testbed in the Open Air Interface (OAI) environment. Thorough experimentation are performed to examine latency and jitter over the link between PHY and MAC. The experimentation setup in this paper follows NGMN recommendations on fronthaul latencies [9].

The remainder of this paper is organized as follows. Section II provides and overview of different C-RAN split solutions and their corresponding impact on the fronthaul. In Section III, we detail our hardware experimental setup for examining the performance of splitting the MAC and PHY layers over Ethernet. Section IV details and analyzes the experimental results in terms of latency and jitter. Finally, section V summarizes the contributions and provides some insight on avenues ahead.

II. RAN SPLIT TOWARDS FLEXIBLE 5G

A. Full Centralization and CPRI

The C-RAN, as one of the enablers of 5G, introduces flexible split of layers in RAN. In the traditional configuration C-RAN consists of RRHs in charge of all radio functions and a BBUs in charge of all the higher layer functionalities, such solution is referred to as “full centralization” in the current literature [18]. Fully centralized C-RAN brings several advantages, such as easier upgrade of network features or expansion of network capacity, simplification of the radio site, load balancing. More importantly, it paves the way for implementing more advance techniques such as Coordinated MultiPoint (CoMP) [19].

Based on the NGMN definition [9], the fronthaul spans distances between the RRH, or radio unit (RU) and the BBU, or central unit (CU). Based on the above traditional configuration, the fronthaul is a point to point link that transports base band radio samples using Common Public Radio Interface (CPRI). In this case, the user data is transmitted in the form of an IQ-data block [20]. The capacity demands for native CPRI transmission is proportional to the evolved Node B (eNB) available bandwidth, the number of active antennas, the quantisation resolution (the number of bits per I or Q sample are 8-20 bits for LTE), the cell load and the user data rates. The Equation (1) shows overall CPRI data rate demands over the fronthaul. Considering a 20 MHz transmission bandwidth, three sectors and four antennas, the overall fronthaul capacity is 14.7 Gbps.

\[
R_{\text{CPRI}} = 2N_{\text{ant}} \times R_s \times N_{(\text{res,CPRI})} \times N_{\text{ovhd}} \times N_{8B/10B}, \quad (1)
\]

where:
- \(N_{\text{ant}}\) corresponds to the number of antennas,
- \(R_s\) is the sampling rate,
- \(N_{(\text{res,CPRI})}\) is the resolution of binary representation of symbols to be transported,
- \(N_{\text{ovhd}}\) is the CPRI overhead,
- \(N_{8B/10B}\) is the overhead due to 8B/10B coding.

The bandwidth requirement becomes even more demanding in 5G, reaching capacities as high as 157.3 Gbps, as estimated by 3GPP [17]. Based on Equation (1), CPRI scales rapidly with both carrier bandwidth and number of antenna elements, which poses an important limiting factor for 5G where massive number of antenna elements are expected to be used. An example of the difference in figures, the required fronthaul data rate for CPRI in both 4G and 5G are compared in Table I. The numbers presented in Table I, for the UpLink (UL) and DownLink (DL) bandwidth requirements, assume deployments of massive multiple input multiple output (Massive MIMO) in the 5G, based on the scenario specified in [17].

To address the extremely high data rate demand of CPRI, number of different techniques proposed in the literature. Compressed CPRI can be used to reduce capacity requirements in places where fronthaul faces bandwidth constraint. In this case, CPRI compression and decompression can enhance utilization of fronthaul link, up to three times [21], [19].

In addition to the throughput demands of CPRI, delay and jitter values must be kept to a minimum. In order to allow a correct operation of layer 2 reliability mechanisms, i.e. HARQ (Hybrid Automatic Repeat Request), and maintain the frame delay variation effect, maximum of 100 \(\mu s\) can be considered as the total fronthaul delay to CPRI transmission.

According to the 3GPP specifications in FDD \(^1\) [22], for a Physical Uplink Shared Channel (PUSCH) transmission\(^2\) in a given sub-frame \(n\), a HARQ acknowledgement feedback is expected in sub-frame \(n + 4\). This is due to the synchronous nature of the HARQ process in the UL, in which the process identification is directly related to the sub-frame number. Hence, if the eNB does not receive the acknowledgement message, the HARQ process will be interleaved, impairing the overall throughput performance [23].

<table>
<thead>
<tr>
<th>Functional split options</th>
<th>Required bandwidth for 4G</th>
<th>Required bandwidth for 5G</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHY-RF (CPRI)</td>
<td>DL: 14.70 Gbps</td>
<td>DL: 157.3 Gbps</td>
</tr>
<tr>
<td></td>
<td>UL: 14.70 Gbps</td>
<td>UL: 157.3 Gbps</td>
</tr>
<tr>
<td>MAC-PHY</td>
<td>DL: 136.9 Mbps</td>
<td>DL: 5.63 Gbps</td>
</tr>
<tr>
<td></td>
<td>UL: 123.2 Mbps</td>
<td>UL: 7.14 Gbps</td>
</tr>
</tbody>
</table>

\(^1\)FDD: Frequency division duplex
\(^2\)PUSCH channel is used for user plane data transmission in the UL
B. Flexible Functional Split

Another approach to consider to eliminate the bandwidth bottleneck of CPRI fronthaul is the use of alternative functional split options between the RU and the CU, whereby less baseband functionalities are centralized. Different functional splits options have been discussed in the literature, and the most common ones are PDCP-RLC (Packet Data Convergence Protocol and Radio Link Control, respectively), MAC-PHY and intra-PHY splits [19].

The fronthaul data rate requirements of a MAC-PHY split is detailed in Table I, based on the numbers from the 3Gpp report [17]. There is a clear reduction of the transmission bandwidth in the MAC-PHY interface, comparing with CPRI, that stems from the nature of the data being transmitted. The MAC-PHY interface transports MAC PDUs instead of IQ-data blocks, as in the CPRI case. Thereby, the first advantage of a MAC-PHY split over CPRI is the lower fronthaul capacity requirement. Moreover, the transport bandwidth varies in proportion to the payload, allowing it to flexibly adapt to the capacity of base stations. Having such an adaptive solution allows the fronthaul to adapt better to the high data rate demands required by 5G. In terms of delay requirements, the round trip transmission deadline is imposed by the MAC layer reliability procedure and will be the same as for the CPRI split.

Using Ethernet networks as a fronthaul network will also introduce framing overhead, and additional challenge to meet the requirements. On the other hand, using Ethernet links allows to:

- Use lower cost-industry standard equipment
- Sharing and convergence with fixed networks
- Enables statistical multiplexing gains when signal has a variable bit rate
- Allows network monitoring and orchestration with the use of virtualization and SDN

Given the capacity demands of CPRI, and the tight synchronisation between both units, the use of Ethernet for this interface can be particularly challenging. Since MAC-PHY layer split has less demanding requirements, both in terms of capacity and synchronisation, Ethernet can be a candid transport network option.

Therefore, in this work we focus on an experimental evaluation of the MAC-PHY functional, using Ethernet as the fronthaul. The main objective is to measure the latency and jitter resulted from data packetization and Ethernet framing.

III. EXPERIMENTAL SETUP

The experimental setup for the evaluation of the fronthaul for MAC-PHY split is depicted in Figure 1. The overall experimental testbed comprises an end to end LTE system from eNB to User Equipment (UE), and all functionalities of the protocol stack are implemented in the eNB as well as in the UE. The communication flow is shown in Figure 1, and a more detailed description is given in the following lines.

The experiment focuses on the DL direction, thus the IP packets are injected in the eNB PDCP layer and are then handled by the whole protocol stack in eNB. In each sub-frame, the RU PHY layer sends an indication to the scheduler function in the CU to prepare the DL data for transmission. The scheduler sends an indication to the PDCP layer which will fetch the IP packets and prepare the PDCP PDUs that are sent to the RLC. The RLC then informs the MAC layer of its buffer occupancy, the MAC/Scheduler function decides how many Bytes to get from the RLC buffer based on the Channel Quality Indicator (CQI) stored for the UE. Once the MAC layer gets the specific number and size of RLC PDUs it composes the MAC PDUs. Afterwards, the data in the MAC layer is packetized to the MAC PDU by adding the MAC-PHY control header (one byte) and the Ethernet header (14 bytes consisting of source MAC address, destination MAC address, and packet type). The Ethernet packet is then transmitted to the RU via Ethernet.

Upon arrival, the RU de-packetizes the Ethernet packet by removing the Ethernet and MAC-PHY control headers, and the PHY layer performs the Cyclic Redundancy Check (CRC) attachment, encoding, scrambling, modulation and Fast Fourier Transform (FFT) functionalities. Then transmits the RF signal to UE, who handles the received DL data from PHY to PDCP in order to extract IP packets. Figure 2 shows a representation of the overall packetization process.

The OAI UE is a fully compliant LTE UE based on the open source software implementation developed by the OAI community [24]. The software runs on a 8 GB RAM with a Xeon 1220, 4 cores server and its connected via USB 3 (Universal Serial Bus 3) to a USRP (Universal Software Radio Peripheral) used to transmit and receive data. The OAI UE is attached to a OAI eNB (fully compliant LTE eNB) where the C-RAN functional split takes place. As depicted in the figure, the OAI eNB is divided in two blocks, the RU corresponds to the OAI PHY block, where all the RF and PHY related functions take place. The second block corresponds to the CU, and it contains all higher layer functionalities (MAC, RLC and PDCP and layer 3 RRC), it runs on an 8 GB RAM with a Xeon 1220, 4 cores server. Both OAI eNB blocks are connected with Ethernet links with capacity of 1 Gb.

For the sake of completeness, Algorithm 1 shows the flow of the experiment running in OAI and the main configuration parameters are listed in Table II. The experimentation is
Algorithm 1 Information Flow Between CU and RU

1: Inputs:
   CU and RU Initial Configuration
2: Run CU and RU
3: eNB PHY → SS, SI
4: Run UE
5: UE Synchronization:
6: UE ← SS, SI
7: UE performs Random Access Procedure
8: if Contention resolution is resolved then
9:   UE moves to Connected mode
10: eNB performs RRC Connection reconfiguration:
11: Establish RAB
12: Download IP Data to UE
13: procedure KPI MEASUREMENTS(Data)
14:   Measure latency and jitter of the fronthaul
15: end procedure
16: end if

TABLE II
OPEN AIR INTERFACE (OAI) PARAMETERS

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carrier Frequency</td>
<td>2.68 GHz</td>
</tr>
<tr>
<td>System Bandwidth</td>
<td>5 MHz</td>
</tr>
<tr>
<td>Frame Type</td>
<td>FDD</td>
</tr>
<tr>
<td>Uplink Tx/Rx Antennas</td>
<td>1 Tx antenna / 1 Rx antenna</td>
</tr>
<tr>
<td>Tx Gain</td>
<td>100</td>
</tr>
<tr>
<td>RX Gain</td>
<td>80</td>
</tr>
</tbody>
</table>

executed in a sequential order, as shown in Algorithm 1. Once UE is connected to the radio access network (RAN), the DL IP packets are injected into PDCP in order to evaluate the fronthaul.

Based on the experimental setup described previously, we run a series of tests to study the feasibility of splitting the MAC and PHY layers with an Ethernet fronthaul network. The main focus of the experimental setup is to study the bandwidth, latency and jitter across the fronthaul network, and assess the suitability of Ethernet for such purpose. We also measure the impact of different packet sizes in throughput, latency and jitter.

A. Analysis of Latency

The latency in this experiment is measured as the round trip time (RTT) of one packet from the MAC to the PHY layer, a graphical representation of the round trip measurement is given in Figure 1. The reason to measure round-trip time rather than a one-way latency is that the former can be measured more accurately in our setup; since the send and receive times are measured at the same physical location, the synchronization between the two machines does not become a precision barrier.

A more detailed explanation of the procedure to measure the RTT is shown in Figure 3. The following events are considered:
1) Network latency to transmit a data request message (RU PHY sending a request to the CU MAC).
2) Software processing time of the request at the CU MAC to prepare data for RU PHY.
3) Packetisation of the prepared data at the CU MAC.
4) Network latency to transmit data from CU MAC to RU PHY.

It is foreseen that transport block sizes will increase in 5G due to the inclusion of higher modulation and coding schemes (MSC). However, in order to support ultra-low latency applications in 5G, such as Tactile Internet [2], small packet sizes are expected. Figure 4 shows the RTT results obtained for different packet sizes tested. There is a slight almost linear increase of latency with the packet size, which is due to the fact that the MAC layer takes more time to prepare the data when a higher packet size is used. Overall the increase in latency is close to 20% from lower to higher packet size, 107.32 µs for 70 Bytes and increases to 128.18 µs for 982 Bytes. Moreover, Figure 5 depicts the probability distribution function (pdf) of the experimental RTT, for the sake of clarity some of the packet sizes have not been included in this figure. Despite the clear difference in average values, there is a consistency all distributions with 1-2 µs deviation from the average value.

Based on the average result, Ethernet can meet the delay requirements of 250 µs which has been agreed so far in the community [9].

B. Jitter Analysis

Apart from the latency performance, jitter is another important limiting factor in any functional split. The jitter is introduced by computation, mainly due to the Operating System (OS) scheduler, which does not always respond in the same manner, even when considering low latency kernels. The transport network as well plays an important role in jitter. In our experimentation setup shown in Figure 1, the data goes through two switches, which will also introduce substantial latency variability and jitter.

The experimental jitter results are shown in Figure 6; average results are almost equal for all packet sizes and close to a 3% of the average latency, however, maximum and minimum values span from 0 to 163 µs, which in principle gives the idea of a high variability of values. Figure 7 shows the pdf of the experimental jitter samples, and it is shown that the probability of such extreme values is very low (close to zero) and the distribution is quite consistent despite the packet size. In particular, all cases analysed show similar average and distribution values’

C. Fronthaul Throughput Calculation

Finally, for each packet size we experimentally calculate the fronthaul throughput. Table III shows throughput and percentage overhead in the fronthaul links, considering the packet description given in figure 2. The percentage of overhead decreases with the Ethernet packet size, as the 15 Bytes overhead is fixed regardless of the Ethernet packet size. The actual Radio Frequency (RF) data rate is less than the fronthaul
data rate since the overhead is included in the calculation of the fronthaul data rate.

V. CONCLUDING REMARKS

In this paper, we examine how using Ethernet as fronthaul can work in C-RAN, with focusing on the MAC and PHY split. In this context, we setup a hardware-based experimentation platform and analyze latency and jitter experience with packetizing data and fronthauling over the Ethernet. We can show that such split is feasible over the Ethernet and has the advantage of not being directly affected by some of the 5G technologies such as massive number of antennas, i.e., massive MIMO. This is indeed a preliminary study and we are expanding our experimentation to investigate various splitting points and different techniques for improving latency and jitter on a given split.

ACKNOWLEDGEMENT

This work has been supported by The Engineering and Physical Sciences Research Council (EPSRC) Industrial Cooperative Awards in Science & Technology (EPSRC CASE) award and by the British Telecom (BT) as well as the Ericsson 5G and Tactile Internet industry grant to Kings College London.

REFERENCES