Cloud-RAN in Support of URLLC

G. Mountaser*, M. Condoluci*, T. Mahmoodi*, M. Dohler*, Ian Mings†
*Centre for Telecommunications Research, King's College London, UK
†British Telecom, Adastral Park, UK

***Abstract*—Flexible network architecture is envisioned as one of the properties of the next generation mobile networks, and is also foreseen to significantly contribute to lowering latency, and addressing Ultra-Reliable Low Latency Communications (URLLC). In this paper, we study the flexibility in Radio Access Network (RAN) configuration in terms of splitting the radio and baseband functionalities between central cloud and distributed entities, and the impact of this flexibility on delivery of URLLC. Three different functionality splits have been implemented in cloud-RAN, and the impact of each of these splits on communication latency and jitter is examined. We modeled traffic based on the 3GPP traffic models and, for completeness, in addition to URLLC, the two cases of massive machine-type communications (mMTC) and enhanced Mobile Broadband (eMBB) models are also implemented. Hence, thorough analyses are performed and recommendation for split point between central cloud and distributed radio units are discussed.**

***Index Terms*—5G; Cloud-RAN; functionality split; URLLC; eMBB; mMTC.**

I. Introduction

The fifth generation (5G) of wireless networks envision innovative radio technologies for ultra dense deployment, improved coverage, higher data rates and lower communication delays to enable the provisioning of novel services [1]. While enhanced mobile broadband (eMBB) and massive machine-type communications (mMTC) can be seen as an extension of services already supported in 4G networks with high data rate and massive connectivity as main requirements, respectively, ultra-reliable low-latency communications (URLLC) represent novel services with unprecedented requirements.

To simultaneously support URLLC, eMBB and mMTC, with their heterogeneous requirements, in a dynamic and on-demand manner, flexible deployment solutions are needed where functionalities can be moved across the network according to service requirements. Such flexibilities could be dynamic placement of network functions across the network through Network Function Virtualization [2], [3] or slicing network to end-to-end separate instances each addressing different requirements [4], [5]. In this direction, flexibility in radio access network (RAN) has been discussed in the context of Cloud-RAN [6], [7], functionalities of RAN are split between central, cloud-based baseband unit and distributed radio units, to achieve advantages such as cooperative solutions, improved load balancing and RAN sharing. The Cloud-RAN architecture offers the possibility to move the baseband unit (BBU) to a central unit (CU) in support of multiple remote radio heads (RRHs). The split could be pre-defined for different network slice, or it can be dynamically changed for different types of traffic or depending on the network conditions, which could be configures via top-level network controller and offered as-a-service–similar studies on different functionality and how it can be offered as-a-service, and controlled by top-level network controller is presented in [8].

An ultimate case of Cloud-RAN, is a fully centralized solution, where all functionality of baseband runs at the central cloud. In this context, the most frequently used standard fronthaul (FH) interface (the link between the CU and the RU) is Common Public Radio Interface (CPRI). The CPRI is a digital interface for encapsulating radio samples between a RU and CU [9]. Such fully centralized architecture, however, requires the support of high bandwidth on FH. Hence other alternatives are studied and eight different options for the split between CU and RU are considered [10], as illustrated here in Fig. 1, and with the details of functionalities in Fig. 1). Having a functionality split also means moving the traffic on the FH from digitized base band signals to IP-based packets, with consequent benefits in terms of cost and traffic multiplexing. The feasibility of MAC-PHY split when considering an Ethernet-based FH has been studied in [11], where the impact of different packetization options has been evaluated. The performance of PDCP-RLC and MAC-PHY functional splits over the LTE protocol stack using copper links has been studied in [12]. In this research work, the performance is measured in term of total achieved throughput for the LTE UE, for the two functional splits using UDP/TCP and SCTP for the transportation of data.

Focusing on URLLC, this paper analyzes how low-latency requirement of this traffic class is met by providing different split between CU and RU. The analysis are performed through an experimental testbed using Software Defined Radio (SDR) and the Open Air Interface (OAI) [14]. The three implemented splits are options 7, intra-PHY, option 6, MAC-PHY, and option 2, PDCP-RLC. The pros and cons of each split in terms of load on the FH for each service under consideration, are then thoroughly studied in this platform.

The remainder of this paper is organized as follows. Section II provides an overview on Cloud-RAN deployment options, discussing full centralization as well as highlighting the main features and characteristics of the three considered splits in this paper. Section III describes the experimental setup as well as traffic models. In Section IV, we present the results achieved from the testbed, in three different splits and under three different traffic models. Finally, conclusive remarks and some future avenues are details in Section V.

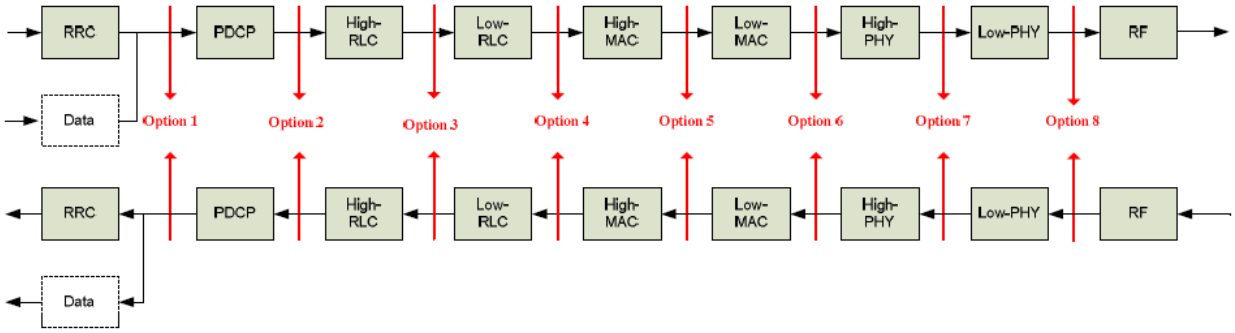


Fig. 1. Options for functional splits for Cloud-RAN and fronthaul [13].

TABLE I
CLOUD-RAN FUNCTIONALITY SPLITS: PROS AND CONS

Split	Pros	Cons
PDCP-RLC	• HARQ is in RU enabling fast retransmission • FH network can handle traffic from different bearers with different priorities • Low OH on the FH	• Only RRC and PDCP are centralized • FH traffic grows with UP/CP traffic load for each bearer
MAC-PHY	• L3 and L2 are centralized • Multiplex multiple bearers into one TB	• HARQ in CU may be challenging to meet HARQ time requirement • OH on the FH depends on PDCP, RLC, and MAC headers and the RB size
Intra-PHY	• Architecture is closed to full centralization • FH load does not depend on the number of bearers and/or UEs	• FH load increases with the bandwidth, number of sectors and antennas • Higher latency as the RU has to receive all packets related to resource elements before starting the IFFT

II. CLOUD-RAN AND VARIOUS LAYER SPLIT

Functional split between BBU and distributed radio units are adopted to address various challenges of radio access networks. As mentioned earlier, the split point for legacy Cloud-RAN is very close to the radio using CPRI. However, eight different options for such split are adopted from the Small Cell Forum document on virtualization functional splits and use cases [15], as illustrated here in Fig. 1. Each of these options will introduce different requirements on the time synchronization and acceptable latency of the FH, but also will introduce different scalability challenges when number of users and amount of data traffic increase. Among all available splits, we will focus our attention on PDCP-RLC, MAC-PHY and Intra-PHY splits (respectively, options 2, 6, and 7 when referring to Fig. 1).

The features of the three splits under consideration in this paper are depicted in more details in Fig. 2 and will be discussed in the remainder of this Section while their pros and cons are summarized in Table I.

A. PDCP-RLC Split

In PDCP-RLC split, RRC and PDCP are executed in the CU while RLC, MAC, PHY and RF are executed in the RU. Fast HARQ can be thus be supported being MAC layer in the RU.

PDCP is responsible for header compression, ciphering, integrity protection and delivering DL processed control and user data to RLC in the form of PDCP PDU as shown in Fig. 2. PDCP delivers PDCP PDUs to RLC once it processes the RRC or IP packets. Since there is no concatenation function in PDCP, if PDCP receives multiple packets from GPRS Tunneling Protocol (GTP), it sends more than one PDCP PDU to RLC via the FH; as a consequence, the FH traffic load increases with control- and user-plane (CP and UP, respectively) traffic. As there is one PDCP entity for each radio bearer, PDCP-RLC split is thus *bearer-based*.

In this approach, it is possible to distinguish between CP and UP traffic. Therefore, the former may be prioritized over the latter in the case of high traffic volume and limited FH capacity. Furthermore, PDCP may use QoS applied to each Radio Access Bearer (RAB) to ensure priority-based treatment of packets on the FH.

B. MAC-PHY Split

In the case of MAC-PHY split, PHY is located in the RU while layer 2 and 3 are centralized. MAC multiplexes MAC

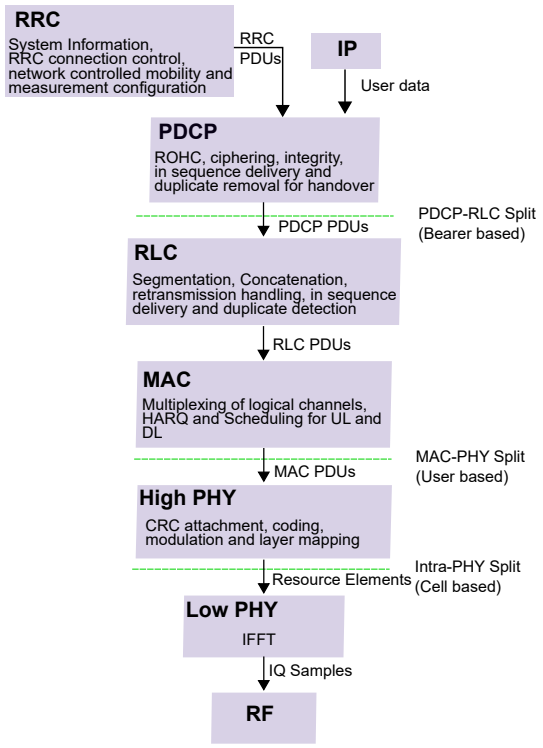


Fig. 2. EUTRAN protocol stack with functions for each protocol layer.

Service Data Units from one or different logical channels onto transport blocks (TBs) then delivers the TBs to PHY. The TB size depends on scheduling decision which considers RLC buffer occupancy, the available bandwidth and selected modulation scheme. As a consequence of MAC multiplexing, the data delivery to PHY via FH is taken per UE and not per radio bearer as in PDCP-RLC split case.

Unlike in PDCP-RLC split where HARQ is located in RU, in MAC-PHY split the HARQ is centralized. Hence this split option is more latency constrained, compared to PDCP-RLC split. In fact, there is a strict requirement of 4 ms HARQ response time set by 3GPP [16]. This requirement will further be restricted in 5G. This aspect may become challenging when considering high-latency or high-loaded FH networks.

C. Intra-PHY Split

In intra-PHY split, PHY functionality is split between CU and RU. The IFFT is performed in RU while other functionalities are performed in CU. Therefore, more functionalities are centralized compared to PDCP-RLC and MAC-PHY splits.

The FH transports resource elements in the usable bandwidth, in which case the capacity of the FH is independent of the actual user traffic. The load on the FH has constant data rate and hence there is no multiplexing gain to be achieved. Since the FH transports resource elements across all the configured OFDM symbols, the FH transmission time granularity can either be symbol or subframe.

In our experimentation, the transmission per symbol is considered to avoid sending too large packet. In this approach, 14 packets are transported in total in each subframe, as there are 14 OFDM symbols in one subframe. In our experimentation, the size of each packet is 12B $\cdot 2 \cdot N_{RB}$, coming from the fact that 1B used to code one sub-carrier (thus, 12B as there are 12 sub-carriers), 2 is because of the I/Q modulation (1B for I and 1B for Q), and N_{RB} represents the number of resource blocks (RBs) the channel bandwidth is composed of.

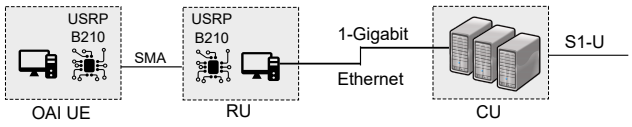


Fig. 3. Setup of the testbed platform.

III. Experimental Setup

Fig. 3 shows the experimental setup to evaluate the performance of Cloud-RAN functionality splits using real time platform OpenAirInterface (OAI) [14].

The experimental testbed consists of a OAI User Equipment (UE) based on the LTE OAI implementation and CU and RU whereby the Evolved NodeB (eNB) functionalities are implemented. The UE runs on a PC (4 GB RAM with an Intel core i5). The CU and RU run on two separate servers (8 GB RAM with a Xeon 1220, 4 cores), connected through an Ethernet link with capacity 1 Gigabit[1]. Two Universal Software Radio Peripherals (USRPs) are used to transmit/receive data between UE and RU. The USRPs are connected to their relevant machines via USB 3.

The experimentation is executed according to the steps shown in Algorithm 1. Radio parameters are listed in Table II. The focus of our experimentation is on downlink (DL) direction. We conducted various experiments to study the impact of the three splits discussed in Section II while running the FH over Ethernet. The functions are shifted between CU and RU according to the functionality split to be evaluated. Once the UE is connected to the RAN and the RAB is established successfully, the IP packets are injected in the CU on top of PDCP.

Traffic pattern for URLLC considers an industry-based closed-loop application [1] and it is modeled with 1000 UEs

[1]In our experimentations, PDCP-RLC and MAC-PHY splits work also when connected through a switch, while the Intra-PHY split does not support this setup due to time constraints. For the sake of uniformity, we thus used a direct Ethernet link to connect the CU and the RU for all the different splits.

TABLE II
Open Air Interface (OAI) parameters

Parameter	Value
Carrier Frequency	2.68 GHz
System Bandwidth	5 MHz (25 RBs)
Uplin Tx/Rx Antennas	1 Tx antenna / 1 Rx antenna
Modulation Schemes	16 QAM (for mMTC) and 64 QAM (for eMBB and URLLC)

Algorithm 1 Information Flow Between CU and RU

Inputs:
    CU and RU Initial Configuration
Run RU and CU
if Connection between CU and RU is established **then**
    eNB PHY → SS,MIB,SI
    Run UE
    procedure CELL SELECTION
        Band scanning
        UE Synchronization:
        UE ← SS,MIB,SI
    end procedure
    procedure RANDOM ACCESS
        UE → preamble
        UE ← Random Access Response (RAR)
        if Contention resolution is resolved **then**
            UE moves to Connected_mode
            procedure RRC CONNECTION RECONFIGURATION
                Establish RAB
            end procedure
            Download IP Data to UE
            procedure KPI MEASUREMENTS(Data)
                Measure latency of the fronnthaul
            end procedure
        end if
    end procedure
end if

receiving packets (e.g., commands for actuators) from the network with a beta distribution (i.e., higher number of UEs to be served at the same time compared to mMTC) within a 10s time interval with packets of 500B [17], [18]. For the sake of completeness, we consider also two additional services as discussed in [1]: *(i)* eMBB, modeled by considering 10 simultaneous active UEs per cell with full buffer bursty traffic FTP model 3 [19] and IP packets of 1500B; *(ii)* mMTC, modeled with 24000 UEs per cell uniformly accessing the network within a time interval of 60s [20] with a packet size of 300B [17] in order to model the reception of an acknowledgement following a report transmission in the uplink.[2]

We used the experimental setup as shown in Fig. 3 to test the different splits discussed in Sec. II for URLLC as well as eMBB and mMTC. In order to evaluate the performance in terms of introduced latency as well as jitter for each split, we performed other series of tests where CU and RU are located at the same machine. This was necessary in order to avoid synchronization issues between two separate machines. In this case, network latency between the functions of the split was emulated by considering the latency figures achieved in the first set of tests, where latency due to physical transmission over Ethernet cable was taken into account.

[2]According to the setting of the services described above, the load does not exceed one IP packet for each Transmission Time Interval (TTI), i.e., 1ms.

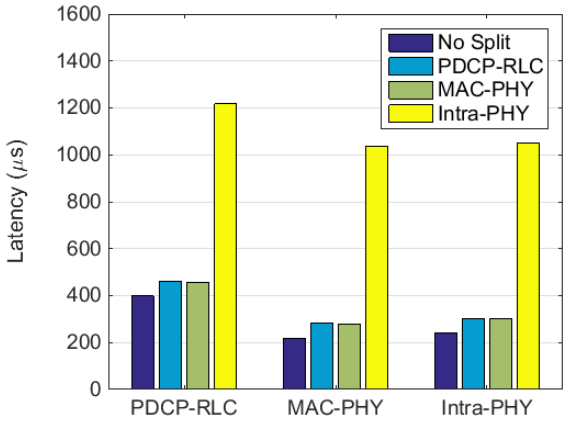


Fig. 4. Latency for an IP Packet from when it is injected to PDCP layer to when it is transmitted to the UE.

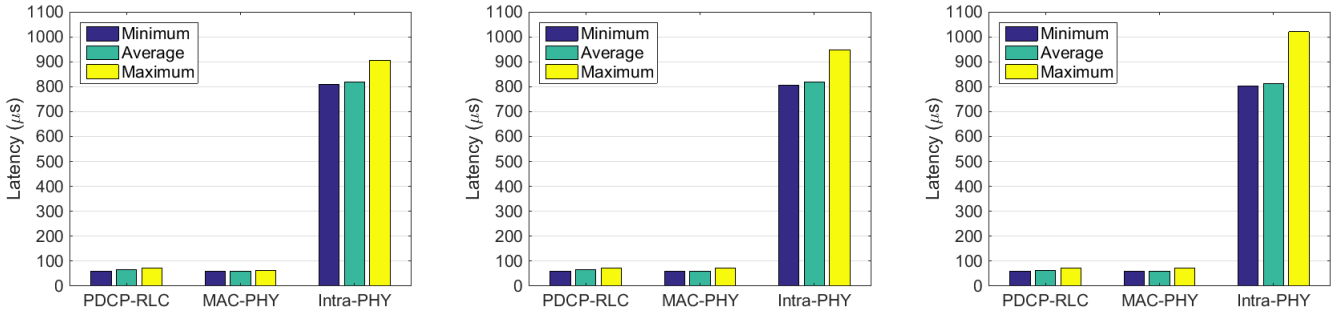
IV. EVALUATION OF FUNCTIONALITY SPLITS FOR DIFFERENT 5G TRAFFIC CLASSES

In this Section, we analyze the performance of the different splits discussed in Sec. II for URLLC and, for the sake of completeness, for eMBB and mMTC.

Fig. 4 analyzes the total latency from PDCP to PHY for UP packets. The aim is to compare the latency introduced by the different functionality splits in addition to the legacy latency introduced by the protocol stack. From this analysis, we can note that the most affected service is eMBB, due to the huge packet size which introduces higher computation load (thus delay). The mMTC and URLLC services are affected by the splits in almost a similar way, with the difference that URLLC has a latency higher of $10\mu s$ than mMTC due to the higher packet size. In this analysis we can see that PDCP-RLC and MAC-PHY splits add a smaller latency than the Intra-PHY split. This is due to the fact that, for the Intra-PHY split, all symbols (i.e., 14 packets) need to be received by the low PHY layer for each transmission time and thus the effective delay introduced has to be considered from when the first packet (related to the first symbol) is transmitted from the CU to when the last packet (related to the last symbol) is received at the RU[3].

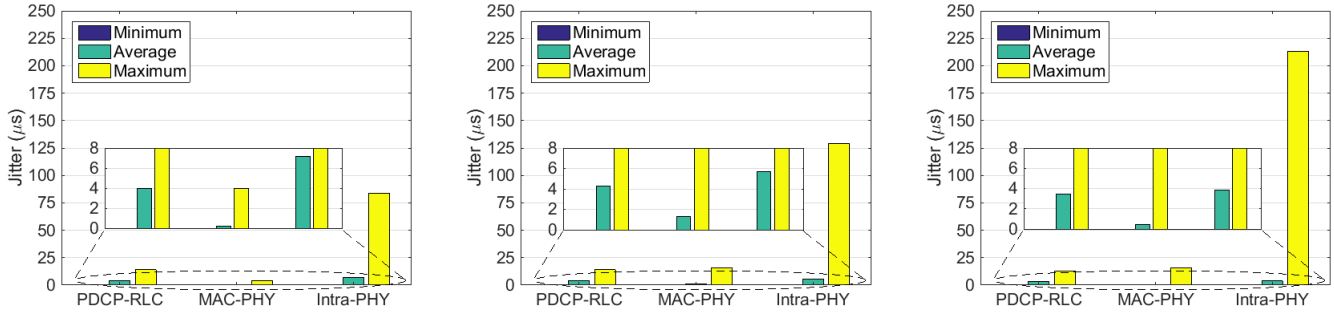
In Fig. 5 the focus is on the latency introduced by each split (i.e., the time interval from when a packet transmission is triggered by the upper layer of the split to when the packet is successfully received by the lower layer of the split). From this analysis, we can note that the PDCP-RLC and MAC-PHY splits work in a more stable way compared to Intra-PHY in terms of added latency. In details, the average latency is almost constant for all the splits and equal to $\sim 65\mu s$ for PDCP-RLC and $\sim 60\mu s$ for MAC-PHY. It is thus interesting to note that the packet size of the different services affect the overall latency

[3]This behavior is due to current OAI implementation, further improvements could be achieved when the RU manages the packets in parallel without waiting for the reception of all 14 packets. The implementation of this feature is out of the scope of this work.



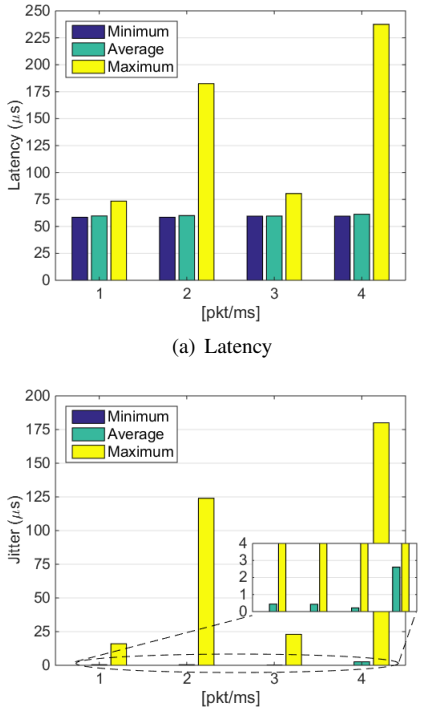
(a) eMBB (b) mMTC (c) URLLC

Fig. 5. Latency from the upper to the lower layer for different splits for 5G services.



(a) eMBB (b) mMTC (c) URLLC

Fig. 6. Jitter for different splits for 5G services (the minimum jitter is equal to 0).



(a) Latency

(b) Jitter

Fig. 7. Latency (a) and jitter (b) for URLLC with MAC-PHY split when varying the number of injected packets.

(as depicted in Fig. 5) but not the one-way latency from upper to lower layers of the split, meaning that the highest source of delay comes from the processing of the packet through the protocol stack. On average, the latency of the Intra-PHY split is equal to $\sim 810 \mu s$. A further analysis can be found in Fig. 6, which depicts the jitter of the different splits. On average, the lowest jitter is guaranteed by the MAC-PHY split, as the MAC and PHY layers work in a synchronous way thus reducing the delay variation. Higher jitter is obtained for the PDCP-RLC split, as in this case the PDCP sends a packet to RLC whenever it receives a packets from upper layers with thus higher latency variation. Finally, the highest jitter is obtained with the Intra-PHY split due to the high number of packets (i.e., 14) transmitted every ms.

After analyzing the performance in terms of latency and jitter, we now focus our attention on the overall pros and cons of the different splits for considered services. For URLLC, MAC-PHY split looks the most adequate solution as analyzed above as it guarantees the lowest and more stable latency. A further analysis is shown in Fig. 7, showing the performance when increasing the number of pkt/ms injected at PDCP layer (as an evaluation of cases with heavy load or other use cases like for instance a mobile gateway simultaneously receives packets to be delivered to non-mobile equipped actuators). Fig. 7 shows that the MAC-PHY split for URLLC has a stable performance when increasing the number of pkt/m, thus

demonstrating the feasibility of this split for URLLC. The MAC-PHY split has in addition the advantage that PDCP is centralized, this being beneficial for multi-RAT convergence to increase reliability.

For delay-tolerant eMBB, all splits may be suitable from a latency point of view, but it is worth reminding that the demand of this service is mainly in terms of data rate. From this point of view, the intra-PHY split does not look to be a suitable solution as the load on the FH directly depends on the available channel bandwidth (in our testbed with 5MHz bandwidth, the FH rate for intra-PHY split is 67.2Mbps for one antenna of one sector as in [6], [21]). The MAC-PHY split is able to aggregate the packets on a UE-basis and the only added OH is in terms of MAC header. This may thus help to reduce the load on the FH in terms of pkt/s, and it thus makes the MAC-PHY a suitable split for delay-tolerant eMBB services. In addition, having a centralization of PDCP and RLC layers would be beneficial for solutions such as multi-RAT convergence, dual-connectivity and better mobility management.

For mMTC, applications dealing with sensing (i.e., delay-tolerant) may be associated to any split, but the intra-PHY split may be a candidate solution for the following reason: all the traffic received by the CU (i.e., both user- and control-plane traffic) will be translated on the FH with a fixed data rate as it depends only on the channel bandwidth. This is beneficial especially for new technologies expected to be used for mMTC such as Narrow-Band IoT (NB-IoT) [22]. By using the same implementation used in our testbed, the overall data rate on the FH for intra-PHY split (where N_{RB} is equal to 1) for one antenna of one sector of NB-IoT would be 2.7Mbps regardless the cell load.

V. Concluding Remarks

In this paper, we study three functionality split options, i.e., PDCP-RLC, MAC-PHY, and intra-PHY, in a cloud-RAN environment, and their impact on delivering URLLC traffic. These three splits are selected out of eight possible options presented by the standardization community such that both lower layer and higher layer splits are examined. Above splits have been implemented in an SDR testbed in the OAI platform with an Ethernet-based FH. For each split, we evaluated latency and jitter. The evaluations show that the MAC-PHY split is the most suitable split option for URLLC as being able to guarantee lowest delay and jitter due to the synchronization needed between MAC and PHY layers.

Future work will be focusing on analyzing the different splits by looking at the overall load (i.e., control plus data traffic) of the FH.

Acknowledgement

This work has been supported by The Engineering and Physical Sciences Research Council (EPSRC) industrial Co-operative Awards in Science & Technology (iCASE) award and by the British Telecom (BT).

References

[1] 5G-PPP, "5G PPP use cases and performance evaluation models," April 2016.

[2] P. Vizzareta, M. Condoluci, C. M. Machuca, T. Mahmoodi, and W. Kellerer, "QoS-driven Function Placement Reducing Expenditures in NFV Deployments," in *IEEE International Conference on Communications (ICC)*, May 2017.

[3] A. Basta, W. Kellerer, M. Hoffmann, K. Hoffmann, and E.-D. Schmidt, "A Virtual SDN-Enabled LTE EPC Architecture: A Case Study for S-/P-Gateways Functions," in *SDN for Future Networks and Services (SDN4FNS)*, pp. 1–7, Nov 2013.

[4] M. Jiang, M. Condoluci, and T. Mahmoodi, "Network slicing management & prioritization in 5G mobile systems," in *European Wireless*, May 2016.

[5] M. Jiang, M. Condoluci, and T. Mahmoodi, "Network Slicing in 5G: an Auction-Based Model," in *IEEE International Conference on Communications (ICC)*, May 2017.

[6] NGMN, "Further Studies on Critical Cloud RAN Technologies." White Paper, March 2015.

[7] 3GPP, "Study on New Radio Access Technology: Radio Access Architecture and Interface (Release 14)," TR 38.801, Aug. 2016.

[8] A. C. Morales, A. Aijaz, and T. Mahmoodi, "Taming Mobility Management Functions in 5G: Handover Functionality as a Service (FaaS)," in *IEEE Globecom Workshops*, December 2015.

[9] "Common Public Radio Interface (CPRI): Interface Specification." V6.0, 2013.

[10] U. Dtsch, M. Doll, H. P. Mayer, F. Schaich, J. Segel, and P. Sehier, "Quantitative Analysis of Split Base Station Processing and Determination of Advantageous Architectures for LTE," *Bell Labs Technical Journal*, vol. 18, pp. 105–128, Jun 2013.

[11] G. Mountaser, M. L. Rosas, T. Mahmoodi, and M. Dohler, "On the Feasibility of MAC and PHY Split in Cloud RAN," in *IEEE Wireless Communications and Networking Conference (WCNC)*, March 2017.

[12] N. Makrisy, P. Basarasy, T. Korakisy, N. Nikaein, and L. Tassiulas, "Experimental evaluation of functional splits for 5G Cloud-RANs," in *IEEE International Conference on Communication (ICC)*, May 2017.

[13] NTT DOCOMO, INC (Rapporteur), 3GPP TSG RAN3, "Study on New Radio Access Technology: Radio Access Architecture and Interfaces." R3-161687, Draft TR 38.801, August 2016.

[14] N. Nikaein, M. K. Marina, S. Manickam, A. Dawson, R. Knopp, and C. Bonne, "OpenAirInterface: A flexible platform for 5G research," *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 5, pp. 33–38, 2014.

[15] Small Cell Forum, "Small cell virtualization functional splits and use cases." Small Cell Forum Document 159.07.02, January 2016.

[16] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA), Physical layer procedures." 36.213, v 11.2. 0, 2013.

[17] "Traffic Model for legacy GPRS MTC." GP 160060, 3GPP GERAN meeting 69, February 2016.

[18] M. Condoluci, M. Dohler, G. Araniti, A. Molinaro, and J. Sachs, "Enhanced Radio Access and Data Transmission Procedures Facilitating Industry-Compliant Machine-Type Communications over LTE-Based 5G Networks," *IEEE Wireless Communications*, vol. 23, pp. 56–63, February 2016.

[19] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects." 3GPP TR 38.814, version 9.2.0, Release 9, March 2017.

[20] 3GPP, "RAN Improvements for Machine-type Communications." 3GPP TR 37.868, version 11.0.0, Release 10, October 2011.

[21] 3GPP TSG RAN WG3, "Transport requirement for CU and DU functional splits options," 2016.

[22] Y. P. E. Wang, X. Lin, A. Adhikary, A. Grovlen, Y. Sui, Y. Blankenship, J. Bergman, and H. S. Razaghi, "A Primer on 3GPP Narrowband Internet of Things," *IEEE Communications Magazine*, vol. 55, pp. 117–123, March 2017.