# Traffic characterization of the MAC-PHY split in 5G networks

Alberto Martínez Alba
*Chair of Communication Networks*
*Technical University of Munich*
Munich, Germany
alberto.martinez-alba@tum.de

Jorge Humberto Gómez Velásquez
*Chair of Communication Networks*
*Technical University of Munich*
Munich, Germany
jorge.gomez@tum.de

Wolfgang Kellerer
*Chair of Communication Networks*
*Technical University of Munich*
Munich, Germany
wolfgang.kellerer@tum.de

***Abstract*—In the 5G radio access network (RAN), the functions of the next-generation eNodeB (gNodeB) are split into a centralized and a distributed unit. Depending on how the split is performed, the amount of traffic generated between the units can be too high to be supported by the current infrastructure. Therefore, a careful characterization of this traffic is needed for every split option. Among the wide array of options, the MAC-PHY split proves to be both promising, as a balanced trade-off between centralization and decentralization, and difficult to characterize, due to the high amount of low level interactions between MAC and PHY layers. Indeed, the MAC-PHY split is frequently considered in the literature as a possible implementation option for the 5G RAN, yet there is no detailed study about the capacity it requires. This paper remedies that by offering a comprehensive analysis of the downlink traffic of a MAC-PHY split for 5G networks. This analysis is backed with both simulative data and measurements from a physical implementation.**

***Index Terms*—MAC-PHY, functional, split, 5G, traffic, fronthaul**

## I. Introduction

The architectural design of the 5G radio access network (RAN) builds upon the idea of centralization. In LTE, all the RAN functions are located at remote sites in order to be close to the radio equipment. This results in high deployment and operating costs, as well as reduced opportunities to implement coordination techniques. To overcome these problems, the C-RAN initative [1] proposes to move all the RAN functions into a centralized location. This decreases costs by reducing the amount of equipment needed and offers the possibility for different base stations (gNodeBs) to coordinate their transmissions. Techniques such as joint transmission, dynamic point selection, or coordinated scheduling are then possible, which improves the quality of experience of the user [2].

Nonetheless, a totally centralized RAN faces a major challenge. Mobile networks are mostly brownfields, in which every component should be reused as much as possible. This implies that the network connecting the new central and remote units (the *fronthaul* network) may have to reuse the infrastructure that formerly connected base stations with their central office (the *backhaul* network). However, C-RAN poses latency and capacity requirements that are often too high to reuse existing backhaul networks [3].

When total centralization is not possible, an alternative proposed by 3GPP is to implement a partially centralized RAN [4], in which a subset of the functions is located at the centralized unit (CU), whereas the rest is deployed at the distributed unit (DU). In a 5G context, these functions are usually the layers of the protocol stack, and the *functional split* is the point of the stack at which the division between CU and DU functions is performed (see Fig. 1). The challenge is to find the split that offers the highest degree of centralization while meeting the limitations of the fronthaul. Thus, the throughput requirements of the selected split and the capacity of the fronthaul network should be accurately matched.

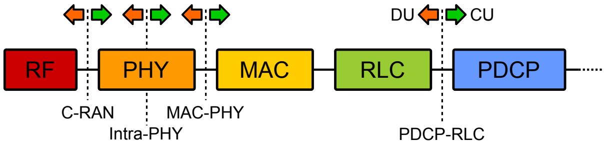


**Fig. 1:** Layers of a 5G RAN and possible functional splits.

For some functional splits, these throughput requirements are simple to predict. For instance, in C-RAN, the fronthaul carries a constant-rate traffic of packetized radio samples, whose magnitude depends only on static parameters, such as the cell bandwidth or the quantization resolution [3]. In PDCP-RLC, the fronthaul traffic is a scaled version of the user data traffic, as the PDCP layer only adds a header to every IP packet. Conversely, the MAC-PHY split introduces a fair amount of complexity into the calculation of the fronthaul traffic, due to the variety of interactions between the MAC and the PHY layers in a 5G RAN.

A good understanding of the fronthaul traffic generated by the MAC-PHY split is motivated by its attractive characteristics, halfway between centralization and decentralization. Indeed, MAC-PHY is the least centralized architecture that still enables coordination techniques, such as coordinated scheduling or coordinated link adaptation, which has been exploited by previous research [5]. In this work, we provide an accurate model of the traffic generated in the MAC-PHY split for a 5G RAN. In summary, our contributions are mainly three: (i) we present an analytical model of the fronthaul traffic generated by the MAC-PHY split, (ii) we provide numerical values for the model, based on the 3GPP specifications, and (iii) we present experimental and simulative results to back our analytical model.

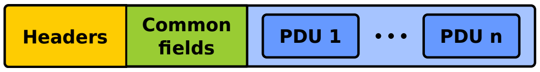


**Fig. 2:** Structure of an nFAPI message.

The rest of the paper is organized as follows. Sec. II summarizes the current literature on the topic. In Sec. III, we introduce nFAPI, an open initiative to implement the MAC-PHY split. Sec. IV explains the notation used throughout the paper. In Sec. V, we derive the traffic of the MAC-PHY split. In Sec. VI, we present experimental results to complement the theoretical analysis. Finally, Sec. VII concludes the paper.

## II. Related work

There are few works directly addressing the capacity requirement of the MAC-PHY split. The main contribution to this area is [3], where the authors present a thorough quantitative study of four functional splits: MAC-PHY, C-RAN, and two different Intra-PHY splits. Although valid as a first approach, their model is rather abstract and does not take into consideration most of the control information that is exchanged between the MAC and PHY layers. Additionally, the 3GPP has addressed the estimation of the MAC-PHY traffic in a technical document [6]. However, this document only includes a rough estimation of the required capacity.

## III. nFAPI specification

In order to model the traffic generated by the MAC-PHY split, we need a protocol that translates the information exchanged by these functions into network packets. To the best of our knowledge, nFAPI [7] is the most complete description of such a protocol. Its goal is to implement all the interactions between MAC and PHY layers as UDP packets, in order that they can be transmitted on an Ethernet network. nFAPI is originally based on the LTE specifications from Rel. 8 to 13, although it can be easily adapted to 5G specifications owing to the similarity between LTE and 5G functions. For this reason, we use the details of this protocol to derive an analytical model for the MAC-PHY split in 5G.

In a nutshell, the operation of nFAPI is as follows. The information exchanged by the MAC and PHY layers is classified into *messages*, which consist of a header region, a common-fields region, and a variable number of packet data units (PDUs) (see Fig. 2). The header region is the same for all messages, and consists of Ethernet, IP, UDP, and nFAPI headers. The sizes of these headers are 14, 20, 8, and 16 B (bytes) respectively, yielding a total size of $s_{\text{H}} = 58$ B. The common-fields region contains auxiliary information that is specific to each message type. In Table 1, we provide the size of this region for the four most common messages. The actual data exchanged by MAC and PHY layers is structured into PDUs. The type and number of these PDUs are variable, as we explain in the following sections. In Table 2, we summarize the size and corresponding message of the most usual PDUs.

| Message | Size of common fields (bytes) |
|---|---|
| DL_CONFIG.request | 13 |
| UL_CONFIG.request | 9 |
| TX.request | 8 |
| HI_DCI0.request | 10 |

**Table 1:** Common fields of the main four nFAPI messages [7].

| PDU name | Size (bytes) | Message |
|---|---|---|
| BCH | 10 | DL_CONFIG.request |
| DLSCH | 79–326 | DL_CONFIG.request |
| DCI | 80 | DL_CONFIG.request |
| ULSCH | 62 | UL_CONFIG.request |
| HI | 18 | HI_DCI0.request |
| DCI0 | 56 | HI_DCI0.request |
| Data PDU | Variable | TX.request |

**Table 2:** Size of the most relevant PDUs in nFAPI [7].

## IV. Notation

Throughout this paper, we use $r_x(t)$ to denote the instantaneous data rate generated by traffic source $x$ at scheduling interval $t$. This data rate is related to the size $s_x(t)$ and the period $\tau_x$ of the corresponding message(s):

$$r_x(t) = \frac{s_x(t)}{\tau_x}. \tag{1}$$

We assume that this expression always holds, thus we skip the definition of $s_x(t)$ when $r_x(t)$ has been already defined, and vice versa. Furthermore, by removing the dependency on $t$, we refer to the mean value over time:

$$r_x \triangleq \mathrm{E}\{r_x(t)\}, \quad s_x \triangleq \mathrm{E}\{s_x(t)\}. \tag{2}$$

In addition, the following relations are also always assumed, without explicit definition:

$$r_x(t) = \sum_{u=1}^{U} r_{x,u}(t), \quad s_x(t) = \sum_{u=1}^{U} s_{x,u}(t), \tag{3}$$

where $r_{x,u}(t)$ is the contribution of user equipment (UE) $u$ to $r_x(t)$ and $U$ is the number of UEs scheduled at interval $t$.

## V. Traffic of the MAC-PHY split

In this section, we provide a detailed description of the information exchanged by the MAC and PHY layers in a 5G RAN. For the sake of brevity, we focus only on the steady-state behavior of the downlink MAC-PHY traffic of a single-carrier cell. Nonetheless, extending the present analysis to the uplink or multiple carriers is straightforward.

In order to predict the downlink traffic in the MAC-PHY split, we need to understand how a data packet is handled by the MAC layer. In a nutshell, the MAC layer divides the user traffic into *data PDUs*, whose size is selected according to the channel quality of each UE. Afterwards, the MAC layer generates one or more *control PDUs* associated to each data PDU, with the intention of assisting the processing of the PHY layer. In the fronthaul, this results in additional traffic that appears every time there is data traffic towards the UEs. Moreover, even if there is no user traffic, the MAC layer periodically generates messages to keep the PHY layer in sync. This creates a constant, user-independent background traffic on the fronthaul. As a result, we can express the fronthaul MAC-PHY traffic $r_{\text{FH}}(t)$ at instant $t$ as the sum of three components:

$$r_{\text{FH}}(t) = r_{\text{BG}}(t) + r_{\text{C}}(t) + r_{\text{U}}(t), \tag{4}$$

where $r_{\text{BG}}(t)$ is the background traffic, $r_{\text{C}}(t)$ is the user-dependent control traffic, and $r_{\text{U}}(t)$ is the user traffic. In the following sections we look into the details of $r_{\text{BG}}(t)$ and $r_{\text{C}}(t)$.

### A. Background traffic

The *background traffic* $r_{\text{BG}}(t)$ has two components: keep-alive messages between MAC and PHY layers and broadcast of system information. Since this traffic is user-independent, we can neglect the influence of time an focus on its average $r_{\text{BG}}$, which can be hence divided into two parts:

$$r_{\text{BG}} = r_{\text{KA}} + r_{\text{SI}}, \tag{5}$$

where $r_{\text{KA}}$ is the traffic resulting from keep-alive messages and $r_{\text{SI}}$ results from the broadcast of system information.

*1) Keep-alive messages:* By design, the PHY layer relies on the decisions of the MAC layer to operate. Hence, the MAC layer sends periodic instructions to the PHY layer, even if there is no user traffic. These instructions can be divided into two different messages that are sent every scheduling interval $\tau_{\text{S}}$: one to configure the downlink (of size $s_{\text{DL}}$), and the other to configure the uplink (of size $s_{\text{UL}}$). Therefore, the data rate of the keep-alive traffic is:

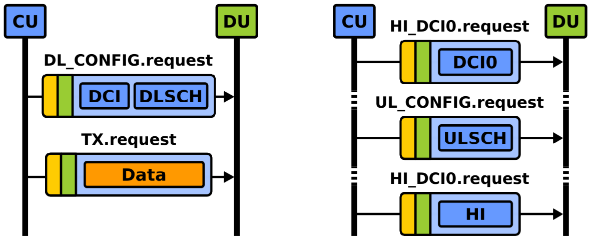
$$r_{\text{KA}} = \frac{s_{\text{DL}} + s_{\text{UL}}}{\tau_{\text{S}}}. \tag{6}$$

In a 5G RAN implementing nFAPI, such messages are DL_CONFIG.request and UL_CONFIG.request, respectively, which are sent every scheduling interval regardless of any user activity. In the absence of user data, both messages consist only of headers and common-fields sections, which contain the frame and subframe numbers, UL/DL configuration, number of control-region sets, and other basic parameters. According to Table 1, their sizes are $s_{\text{DL}} = s_{\text{H}}+13 = 71\,\text{B}$ and $s_{\text{UL}} = s_{\text{H}}+9=67\,\text{B}$. Finally, the scheduling interval ranges from $\tau_{\text{S}} = 62.5\ \mu\text{s}$ to 1 ms, depending on the numerology $\mu \in \{0, ..., 4\}$ [8]. Plugging these values into (6) leads to:

$$r_{\text{KA}} \approx 1.1 \cdot 2^{\mu}\ \text{Mb/s}. \tag{7}$$

*2) System information:* The MAC layer in 5G produces two system information messages to be periodically broadcast by the cell: the Master Information Block (MIB) and the System Information Block 1 (SIB1). The size of the MIB is $s_{\text{MIB}} = 3\,\text{B}$ and it is transmitted every $\tau_{\text{MIB}} = 80$ ms, whereas the size of the SIB1 is approximately of $s_{\text{SIB1}} \approx 18$ B and its period is $\tau_{\text{SIB1}} = 160$ ms [9]. In nFAPI, each MIB entails the transmission of a BCH PDU of size $s_{\text{BCH}} = 10$ B (see Table 2). Conversely, the SIB1 implies transmitting both a DCI and a DLSCH PDU (see Fig. 3a) of sizes $s_{\text{DCI}} = 80\,\text{B}$ and $s_{\text{DLSCH}} = 79\,\text{B}$, respectively. The resulting average traffic $r_{\text{SI}}$ is:

$$r_{\text{SI}} = \frac{s_{\text{MIB}} + s_{\text{BCH}}}{\tau_{\text{MIB}}} + \frac{s_{\text{SIB1}} + s_{\text{DCI}} + s_{\text{DLSCH}}}{\tau_{\text{SIB1}}} \approx 10\ \text{kb/s}. \tag{8}$$



**(a)** DLSCH

**(b)** ULSCH

**Fig. 3:** nFAPI messages sent through the fronthaul associated with data transmission on the DLSCH (Downlink Shared Channel) and ULSCH (Uplink Shared Channel).

This value is negligible compared to the keep-alive traffic, thus we conclude that:

$$r_{\text{BG}} \approx r_{\text{KA}} \approx 1.1 \cdot 2^{\mu}\ \text{Mb/s}. \tag{9}$$

### B. Control Traffic

We define *control traffic* $r_{\text{C}}(t)$ as the traffic on the fronthaul resulting from the transmission of control information from the MAC to the PHY layers. This information is generated only when user data is transmitted or received. Since both downlink and uplink user data streams create control traffic in the downlink, we express $r_{\text{C}}(t)$ as the following sum:

$$r_{\text{C}}(t) = r_{\text{CDL}}(t) + r_{\text{CUL}}(t), \tag{10}$$

where $r_{\text{CDL}}(t)$ is the control traffic due to downlink data transmissions, or simply *downlink control traffic*, and $r_{\text{CUL}}(t)$ is the control traffic resulting from uplink data transmissions, or simply *uplink control traffic*. In the rest of this section, the magnitude of both traffics is explored.

*1) Downlink control traffic:* When a data packet is sent from the gNodeB to a UE, it generates three types of control information along its way. First, the packet is included into PDCP, RLC, and MAC PDUs, and then encapsulated by the nFAPI protocol. Each of these steps adds its own header to the original data packet. Second, the MAC layer generates transmission parameters that the PHY layer uses to transmit the data packet, such as modulation or beamforming vectors. Finally, the MAC layer sends the downlink control information (DCI) for the UE to decode the data packet correctly. If we denote the contribution of these three types of control information by $r_{\text{HD}}(t)$, $r_{\text{L1D}}(t)$, and $r_{\text{DCI}}(t)$, respectively, the following relation holds:

$$r_{\text{CDL}}(t) = r_{\text{HD}}(t) + r_{\text{L1D}}(t) + r_{\text{DCI}}(t). \tag{11}$$

We focus first on $r_{\text{CDL},u}(t)$, the contribution of UE $u$ to the downlink control traffic. Let us define $b_u(t)$ as the content of the downlink data buffer of UE $u$ at time $t$ and the indicator function $I_{b_u}(t)$ such that $I_{b_u}(t) = 1$ if and only if $b_u(t) > 0$. Control information for UE $u$ is only sent if its downlink buffer is not empty, which can be expressed as:

$$r_{\text{CDL},u}(t) = \frac{s_{\text{HD},u}(t) + s_{\text{L1D},u}(t) + s_{\text{DCI},u}(t)}{\tau_{\text{S}}} \cdot I_{b_u}(t). \tag{12}$$

The value of $s_{\text{HD},u}(t)$ should be approximately constant for all UEs, as headers have a fixed size. In contrast, the value of $s_{\text{DCI},u}(t)$ may depend on the size of the DCI format used by the MAC at time $t$, which in turn depends on the system bandwidth and the number of transmission layers supported by the cell. Finally, the value of $s_{\text{L1D},u}(t)$ is a function of the transmission mode, the number of antennas, the bandwidth of the cell, and the number of transmission layers. This results from the transmission of beamforming vectors for non-codebook-based precoding, which need to be provided for every antenna and every *subband* (a fraction of the system bandwidth for which separate channel information is reported).

Let us derive now the average of $r_{\text{CDL},u}(t)$ over a long, stationary period. We can decompose it as:

$$r_{\text{CDL},u} = \hat{r}_{\text{CDL},u} \cdot \eta_u, \tag{13}$$

where

$$\hat{r}_{\text{CDL},u} = \frac{s_{\text{DCI},u} + s_{\text{L1D},u} + s_{\text{HD},u}}{\tau_{\text{S}}} \tag{14}$$

is the maximum possible downlink control traffic generated by UE $u$, and $\eta_u = \Pr\{b_u(t) > 0\}$ is the probability that the buffer is not empty at $t$. It is out of the scope of this paper to provide exact values for this probability, but still general insights can be drawn. Let us consider the average downlink data traffic $r_{\text{U},u}$ experienced by UE $u$ during the interval $T$, which is limited by the capacity of the cell $\hat{r}_{\text{U},u}$ available to UE $u$. It follows that:

$$\eta_u \geq \frac{r_{\text{U},u}}{\hat{r}_{\text{U},u}}, \tag{15}$$

since at least such a fraction of the scheduling intervals need to be used, owing to $s_{\text{U},u}(t) \leq \hat{s}_{\text{U},u}$. This means that if the UE reaches $r_{\text{U},u} = \hat{r}_{\text{U},u}$, for instance when using TCP connections, the maximum downlink control traffic is achieved.

We can get a more accurate description of $\eta_u$ for simple traffic models. For instance, let us consider the case of periodic arrivals of fixed-size packets. That is, we assume that every user packet has a size $s_{\text{U},u}$ and a period $\tau_{\text{U},u}$, yielding a constant throughput of $r_{\text{U},u}$. In addition, we assume that the channel remains approximately constant over the analyzed interval. As a result, we get that every $\tau_{\text{U},u}$, the downlink buffer fills up to $s_{\text{U},u}$ bytes, which takes $\lceil s_{\text{U},u}/\hat{s}_{\text{U},u} \rceil \cdot \tau_{\text{S}}$ to empty. Thus, the average ratio of occupied buffer is:

$$\eta_u = \min\left(\frac{\tau_{\text{S}}}{\tau_{\text{U},u}}\left\lceil\frac{s_{\text{U},u}}{\hat{s}_{\text{U},u}}\right\rceil,1\right) = \min\left(\frac{\tau_{\text{S}}}{\tau_{\text{U},u}}\left\lceil\frac{r_{\text{U}}\tau_{\text{U},u}}{\hat{r}_{\text{U},u}\tau_{\text{S}}}\right\rceil,1\right). \tag{16}$$

From (16), we conclude that if $\tau_{\text{U},u} \leq \tau_{\text{S}}$, then $\eta_u = 1$. This means that if the inter-arrival time of the user packets is lower or equal than the scheduling interval, the maximum control throughput per user is achieved, regardless of the data throughput. This is relevant for transmissions of frequent, small data packets, such as in ultra-reliable low latency communications (URLLC). Conversely, when $\tau_{\text{U},u} \to \infty$, $\eta_u$ reaches the lower bound predicted in (15).

Now that we have an expression for $\eta_u$, we can proceed to calculate $r_{\text{CDL},u}$ as defined in (14). In order to estimate $s_{\text{DCI},u}$, we can do a simple extrapolation. The DCI PDU in nFAPI has a length of 80 B (see Table 2) and contains 49 fields, corresponding to the fields of all the DCI formats used for downlink transmission in LTE. In 5G, there are only 33 different fields in DCI formats 1_0 and 1_1, hence $s^{\text{5G}}_{\text{DCI},u} \approx 54$ bytes. This reduction comes from the removal of deprecated parameters from early LTE releases. The value of $s_{\text{L1D},u}$ is the size of the DLSCH PDU, which depends on the number of antennas $\alpha$, the number of subbands $\sigma$ (directly related to the cell bandwidth), and the number of transport blocks $\beta_{\text{D}}$ transmitted in one scheduling interval. Since the bandwidth-independent part of the DLSCH PDU is due to the DCI parameters, we can extrapolate it from 79 B to 53 B owing to the same reason as before. This results in:

$$s_{\text{L1D},u} = (53 + (3 + 2\alpha)\sigma)\beta_{\text{D}} \approx 4\alpha\sigma\,\text{B}, \tag{17}$$

where the approximated expression holds when the numbers of antennas and subbands are large (such as $\alpha \geq 16$ and $\sigma \geq 15$) and $\beta_{\text{D}} = 2$. Regarding $s_{\text{HD},u}$, for simplicity we assume that the MAC, RLC, and PDCP headers add 10 bytes to the control traffic per UE in one scheduling interval. This is equivalent to the size of two PDCP headers, two RLC headers, and one MAC header, as if two different IP packets of two different bearers were transmitted simultaneously. In addition, we need to consider the header and common fields of the message TX.request that nFAPI employs to transmit user data, thus $s_{\text{HD},u} \approx 12 + \frac{66}{U}$ B. With such values, (14) leads to:

$$r_{\text{CDL}} \approx 32 \cdot 2^{\mu}\alpha\sigma U\bar{\eta} \text{ kb/s}, \tag{18}$$

where $\bar{\eta} \in [\frac{r_{\text{U}}}{\hat{r}_{\text{U}}}, 1]$ is the average $\eta_u$ over all UEs.

*2) Uplink control traffic:* When a data packet is sent from the UE to the gNodeB, the MAC layer of the latter generates three types of downlink control information to assist its transmission and reception. First, the scheduler sends a DCI message to the UE with the resources assigned for uplink transmission. Second, the MAC layer sends uplink parameters that the PHY layer uses to correctly receive the data. Finally, the MAC layer acknowledges the uplink transmission by sending a Hybrid Automatic Repeat Request (HARQ) indicator to the UE. If we denote the contribution of these three types of control information by $r_{\text{DCIU}}(t)$, $r_{\text{L1U}}(t)$, and $r_{\text{HI}}(t)$, respectively, the following relation holds:

$$r_{\text{CUL}}(t) = r_{\text{DCIU}}(t) + r_{\text{L1U}}(t) + r_{\text{HI}}(t) \tag{19}$$

As in the downlink case, the average of $r_{\text{CUL},u}(t)$ over a long, stationary period can be expressed as:

$$r_{\text{CUL},u} = \hat{r}_{\text{CUL},u} \cdot \gamma_u, \tag{20}$$

where

$$\hat{r}_{\text{CUL},u} = \frac{s_{\text{DCIU},u} + s_{\text{L1U},u} + s_{\text{HI},u}}{\tau_{\text{S}}} \tag{21}$$

is the uplink control traffic generated by UE $u$ under the full-buffer assumption, and $\gamma_u = \Pr\{v_u(t) > 0\}$ is the probability that the content of the uplink buffer $v_u(t)$ for UE $u$ is not empty at time $t$. The same conclusions derived for $\eta_u$ can be applied to $\gamma_u$ for uplink traffic.

Regarding the components of $\hat{r}_{\text{CUL},u}$, in nFAPI the DCI and the HARQ acknowledgments are sent as DCI0 and HI PDUs, respectively, within a message HI_DCI0.request, as depicted in Fig. 3b. The sizes of both PDUs are shown in Table 2. We can assume that the size of the header and the common-fields region of HI_DCI0.request (see Table 1) is split between the two PDUs, as there will be the same number of DCI0 and HI PDUs on average. The size of the DCI for uplink scheduling has been reduced in 5G with respect to 4G, containing only 24 fields (combining formats 0_0 and 0_1), whereas in LTE there are 34 fields (in formats 0 and 4). After extrapolating, we obtain $s_{\text{DCIU},u} \approx 40 + \frac{34}{U}$. The HARQ acknowledgments to uplink transmissions are asynchronous and their delay is not fixed. This translates into more control information to be sent than that shown in Table 1, which according to our measurements is $s^{\text{5G}}_{\text{HI},u} \approx 30+\frac{34}{U}$ bytes. Regarding the physical layer parameters for uplink reception, they are sent as one or two ULSCH PDUs within the message UL_CONFIG.request, thus $s_{\text{L1U},u} = 62\,\beta_{\text{U}}$ B, where $\beta_{\text{U}}$ is the average number of uplink transport blocks received simultaneously. In contrast to the downlink case, the number of uplink physical parameters does not scale with the number of antennas or subbands, as only full-bandwidth codebook-based precoding is used in the uplink, hence only the index of the codebook is sent. In total, the uplink control traffic is:

$$r_{\text{CUL}} \approx (0.54 + 1.58U)\cdot 2^{\mu}\bar{\gamma}\ \text{Mb/s} \tag{22}$$

where $\bar{\gamma}$ is the average $\gamma_u$ over all UEs.

### C. Average fronthaul traffic

At this point, we can combine expressions (7), (8), (18), and (22) to obtain a compact estimation of the average fronthaul traffic $r_{\text{FH}}$ of a 5G RAN implementing a MAC-PHY split:

$$r_{\text{FH}} \approx 2^{\mu}\cdot\left[1100+32\alpha\sigma\bar{\eta}U+(540+1580U)\bar{\gamma}\right]+r_{\text{U}}\ \text{kb/s}. \tag{23}$$

Equation (23) reflects the linear dependency of the overhead traffic on the number of antennas $\alpha$, the number of subbands $\sigma$, and the frequency of the user data packets $\bar{\eta}$ and $\bar{\gamma}$. In order to grasp the magnitude of the overhead traffic, let us consider, for example, a 200 MHz carrier with two bandwidth parts ($\sigma = 36$), $\alpha = 128$ antennas, $U = 10$ simultaneously scheduled users [10], and $\bar{\eta} = \bar{\gamma} = 1$. With those values, (23) leads to:

$$r_{\text{FH}} \approx 1.5\cdot 2^{\mu}+r_{\text{U}}\ \text{Gb/s}. \tag{24}$$

We can see that the average overhead traffic $\Delta r_{\text{FH}} = r_{\text{FH}} - r_{\text{U}}$ ranges from $\Delta r_{\text{FH}} = 1.5$ to 24 Gb/s, depending on the numerology, which is actually comparable to the capacity the air interface. This estimation of the overhead traffic is significantly higher than that presented by previous research [3] [6].

## VI. Experimental Results

The accuracy of the theoretical model presented in the previous section is put to the test in two different ways. On the one hand, we employ a real testbed of a RAN featuring a MAC-PHY split. On the other hand, for those cases in which the testbed cannot not be used, a simulator is used. In order to produce experimental results as accurate as possible, in both strategies actual nFAPI packets are generated in real time between MAC and PHY functions.

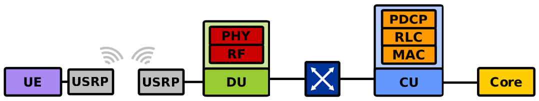


**Fig. 4:** LTE testbed implementing a MAC-PHY split.

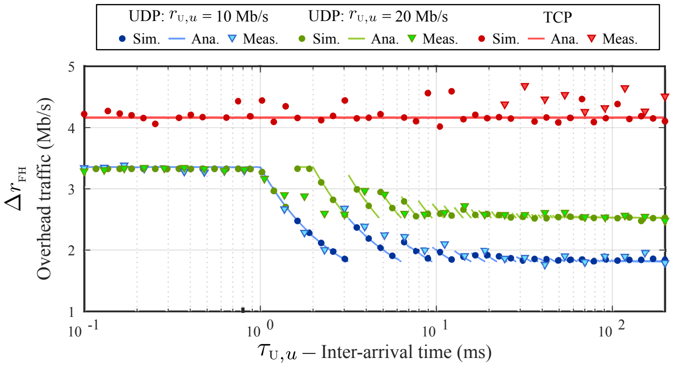


**Fig. 5:** Overhead traffic in MAC-PHY as a function of the inter-arrival time of the received packets (for UDP) or bursts (for TCP).

The testbed consists of four Intel i7 PCs (for the UE, DU, CU, and core) and two USRP B210 SDRs (for the UE and DU), as depicted in Fig. 4. This testbed implements a 10 MHz single-antenna LTE cell ($\alpha = 1$, $\sigma = 9$, and $\mu = 0$), and achieves a maximum downlink throughput of $\hat{r}_{\text{U}} = 31.7$ Mb/s. We use srsLTE [11] as the software platform, with custom modifications to implement the nFAPI MAC-PHY split. For the simulations, we employ an extended version of the open-nFAPI simulator [12] to enable 5G characteristics (such as high number of antennas and subbands).

We present three experiments regarding the magnitude of overhead traffic $\Delta r_{\text{FH}}$. The first experiment measures the impact of the inter-arrival time of the user packets on the total fronthaul traffic. This impact is modeled by the variables $\bar{\eta}$ and $\eta_u$ as defined in (16). We generate constant-rate UDP downlink streams of $r_{\text{U},u} = 10$ Mb/s and 20 Mb/s, as well as TCP streams. For the UDP streams, we vary the transmission period of the datagrams from $\tau_{\text{U},u} = 0.1$ ms to 200 ms. For the TCP streams, an artificial delay is added to the fronthaul link of the simulator and the testbed to modify the round-trip time (RTT) of the network and hence the time between bursts. This is done from $\tau_{\text{U},u} = 0.1$ ms to 200 ms in the simulations, and from $\tau_{\text{U},u} = 20$ ms to 200 ms in the testbed experiments, as 20 ms is the actual round-trip time. After feeding both the testbed and the simulator with every stream for at least 20 seconds, the average overhead traffic is measured. The results are shown in Fig. 5, where we can see the measured, simulated, and predicted overhead traffic. For the UDP streams, we see that the model matches very closely the measurements and simulation results. It therefore confirms that user traffic with low inter-arrival packet times generates the maximum control overhead regardless of its average throughput, as foreseen in (16). In contrast, when the inter-arrival packet time increases, the model predicts a discontinuous decrease down to an overhead traffic of 2.55 Mb/s for $r_{\text{U},u} = 20$ Mb/s, and 1.85 Mb/s for $r_{\text{U},u} = 10$ Mb/s,

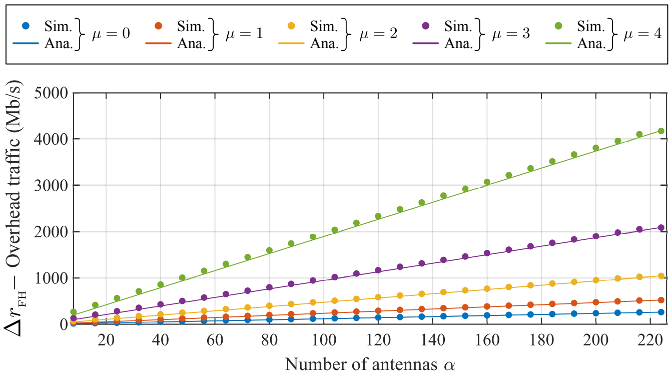


**Fig. 6:** Overhead traffic in MAC-PHY as a function of the number of antennas for a 200 MHz carrier with two bandwidth parts ($\sigma = 36$), $U = 1$, and $\mu \in \{0, \ldots, 4\}$.

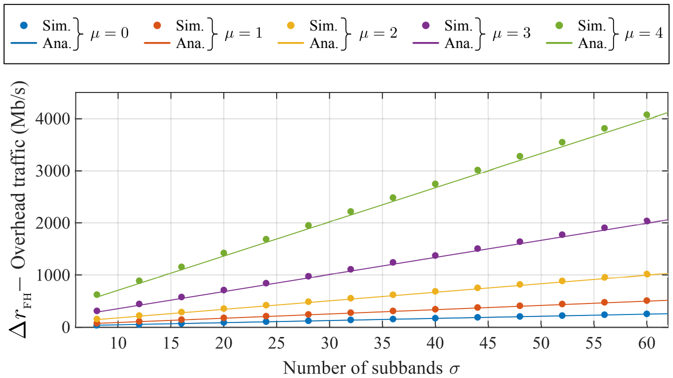


**Fig. 7:** Overhead traffic in MAC-PHY as a function of the number of subbands for $\alpha = 128$ antennas, $U = 1$, and $\mu \in \{0, \ldots, 4\}$.

which are exactly the observed values. For the TCP streams, we confirm that the RTT of the network does not have an impact on the control traffic that is generated. We also observe that this traffic is higher that in the UDP case, because of the uplink transmissions of ACKs.

The following two experiments are simulations, in which both the uplink and downlink buffer are always full to guarantee maximum overhead. The second experiment shows the impact of the number of antennas $\alpha$ on the overhead traffic $\Delta r_{\text{FH}}$ for a 200 MHz carrier with two bandwidth parts ($\sigma = 36$), $U = 1$, and $\mu \in \{0, \ldots, 4\}$. The results are shown in Fig. 6, along with the predicted values. We can see that the overhead traffic grows linearly with the number of antennas and closely resembles the behavior predicted by (23). We observe that, even for a single scheduled user, the overhead traffic of the MAC-PHY split may surpass 1 Gb/s when the number of antennas is large (as in massive MIMO).

Finally, the last experiment shows the impact of the number of subbands (related to the carrier bandwidth) on the overhead traffic produced by a single UE ($U = 1$) with $\alpha = 128$ antennas. The results of are shown in Fig. 7. We observe that the overhead traffic also grows linearly with the number of subbands, as predicted by (23), and may reach values of several Gb/s when the carrier bandwidth is large.

## VII. Conclusion

The next-generation RAN features the division of its functions into centralized and distributed units. Among the options to perform this division, the MAC-PHY split offers a good compromise between the benefits of centralization and decentralization, but the fronthaul traffic that it generates is difficult to estimate. In this paper, we present a complete characterization of such traffic for 5G networks. We base on open specifications to derive an analytical model, and we complement it with measurements and simulations. We find that the MAC-PHY split generates an amount of overhead traffic on the downlink comparable that has been underestimated by previous research. We also conclude that this traffic heavily depends on the number of antennas, system bandwidth, and user-traffic characteristics.

## Acknowledgment

This work is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 647158 - FlexNets). The authors alone are responsible for the content of the paper.

## References

[1] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud ran for mobile networks - a technology overview," *IEEE Communications surveys & tutorials*, vol. 17, no. 1, pp. 405–426, 2015.

[2] A. Martínez Alba, A. Basta, J. H. Gómez Velásquez, and W. Kellerer, "A realistic coordinated scheduling scheme for the next-generation RAN," in *2018 IEEE Global Communications Conference: Next-Generation Networking and Internet (Globecom2018 NGNI)*, Dec. 2018.

[3] U. Dötsch, 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, no. 1, pp. 105–128, 2013.

[4] A. Maeder, M. Lalam, A. De Domenico, E. Pateromichelakis, D. Wubben, J. Bartelt, R. Fritzsche, and P. Rost, "Towards a flexible functional split for cloud-ran networks," in *Networks and Communications (EuCNC), 2014 European Conference on*. IEEE, 2014, pp. 1–5.

[5] A. Martínez Alba, J. H. Gómez Velásquez, and W. Kellerer, "An adaptive functional split in 5g networks," in *2019 IEEE INFOCOM WKSHPS-3rd Workshop on Flexible and Agile Networks: 5G and Beyond (FlexNets*, 2019.

[6] 3GPP, "Transport requirement for CU-DU functional splits options," 3rd Generation Partnership Project (3GPP), TDoc R3-161813, 08 2016.

[7] Small Cell Forum, "FAPI and nFAPI specifications," Document 082.09.05, 05 2017, release 9.0.

[8] 3GPP, "NR; Physical channels and modulation," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.211, 09 2018, version 15.3.0.

[9] ——, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Packet Core (EPC); Common test environments for User Equipment (UE) conformance testing," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 36.508, 09 2018, version 15.3.0.

[10] O. Iosif and I. Banica, "On the analysis of packet scheduling in downlink 3gpp lte system," *CTRQ*, vol. 106, 2011.

[11] I. Gomez-Miguelez, A. Garcia-Saavedra, P. D. Sutton, P. Serrano, C. Cano, and D. J. Leith, "srslte: an open-source platform for lte evolution and experimentation," in *Proceedings of the Tenth ACM International Workshop on Wireless Network Testbeds, Experimental Evaluation, and Characterization*. ACM, 2016, pp. 25–32.

[12] "Open-nfapi," https://github.com/cisco/open-nFAPI.