Delay-Aware Green Hybrid CRAN

†Abdulrahman Alabbasi, †Cicek Cavdar

†Communication Systems Department, KTH Royal Institute of Technology, Sweden
Email: †{alabbasi, cavdar}@kth.se

***Abstract*—As a potential candidate architecture for 5G systems, cloud radio access network (CRAN) enhances the system's capacity by centralizing the processing and coordination at the central cloud. However, this centralization imposes stringent bandwidth and delay requirements on the fronthaul segment of the network that connects the centralized baseband processing units (BBUs) to the radio units (RUs). Hence, hybrid CRAN is proposed to alleviate the fronthaul bandwidth requirement. The concept of hybrid CRAN supports the proposal of splitting/virtualizing the BBU functions processing between the central cloud (central office that has large processing capacity and efficiency) and the edge cloud (an aggregation node which is closer to the user, but usually has less efficiency in processing). In our previous work, we have studied the impact of different split points on the system's energy and fronthaul bandwidth consumption. In this study, we analyze the delay performance of the end user's request. We propose an end-to-end (from the central cloud to the end user) delay model (per user's request) for different function split points. In this model, different delay requirements enforce different function splits, hence affect the system's energy consumption. Therefore, we propose several research directions to incorporate the proposed delay model in the problem of minimizing energy and bandwidth consumption in the network. We found that the required function split decision, to achieve minimum delay, is significantly affected by the processing power efficiency ratio between processing units of edge cloud and central cloud. High processing efficiency ratio (≈ 1) leads to significant delay improvement when processing more base band functions at the edge cloud.**

***Index Terms*—5G, network architecture, cloud RAN, end-to-end delay, network function split, virtualized cloud RAN.**

I. Introduction

The huge demand on higher capacity motivated the research on ultra-dense radio networks, which requires a huge/centralized processing that can be realized by CRAN architecture. However, CRAN imposes high bandwidth requirement on the fronthaul link. Hence, hybrid cloud radio access network (H-CRAN) is proposed to overcome this strict requirement. In addition to sharing the processing and virtualization of functions, H-CRAN provides a multiplexing gain for saving system's energy and cost. In our previous work, in [1], we have proposed and solved an optimization problem that formulates the impact of different function processing splits on the interplay between energy and bandwidth consumption in H-CRAN. Our proposed architecture contains a central cloud, i.e., which has huge computational capabilities and efficient power consumption, connected via fiber to edge clouds, i.e,. aggregation nodes which are located closer to the user equipment and can conduct base band processing (with lower efficiency). In this work, we propose an end-to-end delay model, which reflects the impact of centralizing and/or distributing the communication functions processing at Central-Cloud (CC) and/or Edge-Cloud (EC).

Several works have studied the delay and synchronization performance of CRAN. An early results on the field trials of CRAN's delay (with(out) the Wavelength Division Multiplexing (WDM) optical ring) are reported in [2]. The authors of [3] looked at reusing existing packet-based network (e.g. Ethernet) to possibly decrease deployment costs of fronthaul of CRAN and cost of Baseband Unit (BBU) resources. Accurate phase and frequency synchronization imposes a challenge in packet-based fronthaul. They verified the feasibility of using the IEEE 1588v2, known as Precision Time Protocol (PTP), for providing accurate phase and frequency synchronization in the fronthaul. In [4], the authors have proposed a novel scheme to reduce the latency of a CRAN architecture. In a separated data and control planes architecture, they proposed a user-centric decision on whether to retransmit or not based on some simple feedback from the RU. In CRAN settings, authors of [5] have proposed a queue-aware power and rate allocation for delay-sensitive traffic and formulate it as a Markov decision process. On the other hand, the authors of [6] have evaluated the impact of function splits on the energy and cost savings of the CRAN network. Note that, unlike our work, none of the above literatures has considered the impact of variable function processing splits on the delay performance of end users' requests.

In this work, we utilize our previously proposed architecture of H-CRAN with two transportation links, i.e., midhaul (connects CC-to-ECs) and fronthaul (connects ECs-to-RUs) [1]. In the H-CRAN architecture, digital units (DUs) are deployed at both ECs and CC to allow processing of baseband functions in each or both of the clouds, EC and/or CC. Also, H-CRAN architecture enables the shared digital units at CC to process different users' and cells' functions which leads to higher energy saving gain, hence it is a green architecture. We propose an end-to-end delay model per user's request that utilizes the H-CRAN architecture and the function split model to evaluate the delay performance of each individual content request. Our delay model considers all the delays induced by the network components from CC to EC to RU, then to user equipment (UE). We also propose three interesting technical research directions to incorporate our function split based delay model in several optimization problems, e.g, energy and midhaul's bandwidth minimization problems. Finally, we evaluate a use-case of one user delay model based on all possible function processing splits at EC and CC.

The organization of the remaining sections is as follows. Section II recalls the architecture of H-CRAN, and presents the functional split model. Section III describes the function split oriented delay model. Section IV presents technical thoughts about incorporating the proposed delay model in energy related optimization problems. Finally, Sec. V presents

978-3-901882-90-6/17 ©2017 IFIP

the result of a use-case, which evaluates the delay performance of a one user based on function processing split among the dual-sites.

II. Network Architecture

In this section, we present the system/network architecture, while highlighting the distinctions against reference architectures, and function splits options, similar to [1]. The introduction of these architectures will pave the way for understanding the proposed delay model, later in Sec. III.

A. H-CRAN Architecture

We present a hybrid architecture that employs dual-site processing in H-CRAN. In this architecture, DUs are deployed at both CC and EC, so that baseband processing can be flexibly provisioned by a chain of virtualized functions for a RU or for a UE, while the associated traffic is transported through the midhaul or fronthaul links in the network. We call this architecture as hybrid cloud radio access network (H-CRAN), as shown in Fig. 1.

H-CRAN is a three-layer architecture, which consists of cell layer (the coverage of RU is referred to as a 'cell'), Edge-Cloud (EC) layer, and Central-Cloud (CC) layer. Cell layer consists of cells that are being densified, each serving several UEs. A group of cells are connected to a EC as an aggregation point. The fronthaul between a cell and a EC can be implemented using a short fiber (as in conventional settings), or wireless links, e.g. mmWave links [7] or free-space optical links [8][1]. The ECs is connected to CC via midhaul using various technologies from expensive dark fiber solutions to cost-efficient passive optical network (PON) families or other Ethernet-based technologies. The midhaul technology considered in this study is time and wave division multiplexing (TWDM)-PON [9], and each midhaul link is a wavelength channel, which needs an optical network unit (ONU) at EC and a Line-Card (LC) at CC as transceivers. We assume that there are optical switches at the data-center (CC), area node (EC), and access node (not used in case of using Milli-Meter wave (m-Wave) for fronthaul link)[2]. We also assume that there is an Ethernet switch at the CC.

Edge cloud layer and central cloud layer contain DUs. These DUs are able to accommodate and process virtualized functions of the requested contents and network processes. Hence, the DUs are capable of sharing their computational resources by any connected RUs (if implemented in general purpose servers). For example, in upstream, traffic from cells can be partially processed at edge cloud so that bandwidth requirement can be relaxed for midhaul, then remaining processing will be conducted at central cloud. However, EC is usually less energy-efficient than CC, because the number of DUs, associated with RUs, at the CC is larger than that in each EC. Hence, sharing infrastructure equipment results in higher energy saving at CC. The trade-off becomes whether to distribute functions at EC (to save midhaul bandwidth and improve the delay), or to centralize more functions at CC in central cloud layer (to save power).

[1]Our architecture is not hard-wired to any specific fronthaul technology within EC, as they are not the focus of our study.

[2]These optical switches exist for several metro optical network topologies, e.g., ring, which are used to provide fiber connections in cities.

B. Reference Architectures

We define two extreme cases with no functional split as reference architectures for the performance analysis.

1) Edge-CRAN where all the baseband functions are centralized at the edge cloud and the connection to the central cloud is provided by a backhaul. In this case, DUs are stacked at a nearby cabinet within the EC. DUs, and infrastructure in cabinet, are dedicatedly serving RU of the base station (BS), and cannot be shared by other BSs, which leads to low energy efficiency. Since baseband processing is fully conducted at EC, the conventional backhaul requires a small amount of bandwidth as perceived by UEs.
2) Central-CRAN where all the baseband functions are centralized at the central cloud. In this case, sharing infrastructure for the required baseband processing results in reducing the power consumption [10]. However fronthaul must be considerably prolonged by using dedicated fibers, optical transport network, or PON [11] because DUs are located at CC. Since no baseband processing is conducted at EC, large delay and bandwidth consumption are imposed on the system performance.

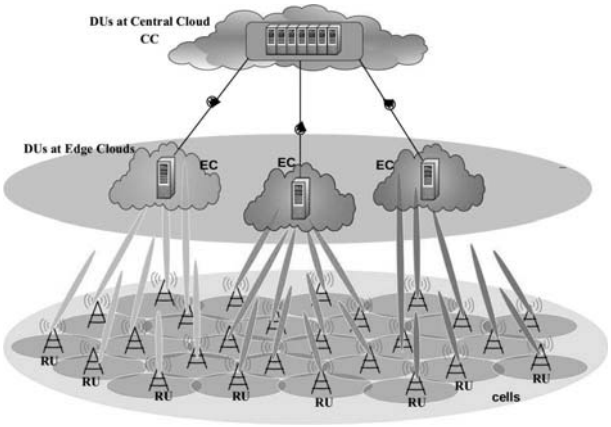

Fig. 1. H-CRAN architecture [1].

C. Functions Split Model

To study the function distribution against function centralization, we model the functional split of baseband processing chain for cells and users, as shown in Fig. 2. First, baseband processing for a cell and its users is modeled as a chain of functions, which includes m Cell-Processing (CP) functions and n User-Processing (UP) functions[3]. CPs are a sequence of functions in physical layer that are dedicated for processing signals from a cell, when signals of UEs are multiplexed. For example, in upstream, CPs includes: (1) serial-to-parallel conversion and common public radio interface (CPRI) encoding, (2) pre-distortion, filtering, up/down sampling, and Time-Domain estimation, (3) fast fourier transform and its

[3]In this study, $m = 4$ and $n = 4$, as described in [12]–[14].

related operations,(4) Resource mapping, etc. The per-cell processing will be terminated at CP_m, and the signals from a cell will be de-multiplexed as multiple signal streams, each belonging to a UE. Then, UPs is a sequence of functions that will continue to process the signal streams on a per-UE basis[4], including (1) equalization, inverse discrete Fourier transform, (2) modulation/demodulation MIMO (de)mapping and (pre)coding, (3) forward error correction, turbo decoding, and (4) upper layers functions.

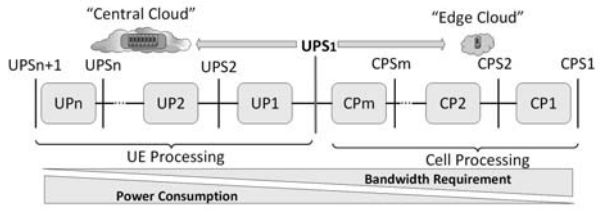

Fig. 2. Function split model [1].

As shown in Fig. 2, functional split can happen before CP_1, after UP_n, or between any two functions. Note that CP split 1 (CPS_1) is the initial attempt to implement CRAN, which is based on full baseband centralization. UP split n+1 (UPS_{n+1}) is implemented by Edge-CRAN, characterized by a fully distributed deployment.

III. Delay Model Based on Function Split

In this section, we present the proposed end-to-end delay model per user's request. Figure 3 briefly describes the system's components that contribute to the total end-to-end delay. That is, all network components in CC-to-EC segment[5], EC-to-RU segment[6], and RU-to-UE segment[7]. This delay model is related to the distribution or/and centralization of the communication functions processing. Different function processing splits result in content processing at either CC, or EC, or partially at both of them, hence, contribute differently to the total systems delay. In the following, we briefly describe how each component contribute to the overall systems delay.

If the functions processing split decision resulted in partial/full processing at CC, it will induce the following delay components:

- Accumulative Delay induced by communication functions processing at CC. This processing is conducted based on each radio sub-frame. The amount of processing is related to the decided split point.
- Accumulative delay induced by encapsulating the data resulted from different splits into several optical frames, each has the delay of 125usec. Note that in our settings, we assume that the user's requested content needs N_{rsf} radio sub-frames to be delivered within D^u_{thr} delay threshold accepted by user u. The function processing delay is evaluated individually for each radio sub-frame, then accumulated for all required radio sub-frames. Each split, will result in a larger amount of data (than the requested one) which needs N_{of} optical frame to transport this data via midhaul connection.
- Constant delay induced by optical conversion and processing devices[8], the optical propagation.
- Constant delay induced by optical switch at the data-center and the area nodes[9], in addition to the Ethernet switch at the data-center.

When all functions processing are conducted at EC, we assume that no delay is induced from CC. Because the optical frame size is enough to support the requested data, i.e., $N_{of} \approx 1$.

Furthermore, if the function processing split decision resulted in full/partial processing at EC, it will induce the following delay components:

- Accumulative Delay induced by communication functions processing at EC. This processing is conducted based on each radio sub-frame. The amount of processing is related to the decided split point.
- Accumulative delay induced by encapsulating the requested content over multiple N_{RSF} radio sub-frames, and push it to the m-Wave link (fronthaul) which afterward will be pushed to the radio frequency-link to the end user without accumulating more delay (Unless different frame and protocol are used at fronthaul, which is not the case in this model).
- Constant delay induced by summing m-Wave converting delay and m-Wave propagation delay.

Note that the aforementioned delay, at EC, will always contribute to the total delay, except the processing part (when all functions processing are conducted at the central cloud). The last mile delay (from RU to UE) is constant, it is induced from the m-Wave conversion delay, access node delay (in-case the fronthaul link is based on fiber), the radio frequency (RF) propagation delay, and the user processing delay.

The calculation of the delay induced by communication functions processing depends on two major factors, one is the giga operation per second (GOPS) required per function processing (referenced by unit time) and the other is the processing power of the equipment. Table II (in Appendix A) lists, in the first column, all the necessary digital sub-components, which contribute to the overall delay [14]. In the second column, the associated GOPS per each communication sub-component function is listed. The exponent factors, that shows the impact of using parameters that differ from the reference case (i.e., SISO, 20 MHz, 64-QAM, coding rate 1), are listed in the third to the seventh columns. Utilizing the

[4]Note that we allocate fixed number of resource blocks (RBs) for each UE such that at full load cell the assigned 20 MHz per cell is enough to serve all users per cell.

[5]The components which contribute to CC-to-EC delay segment are: DUs at CC, data-center's Ethernet switch, number of required optical frames per split, optical switch at the data-center, optical conversion between electric and optical signal including all processing needed for optical transmission, and the optical propagation delay.

[6]The components which contribute to EC-to-RU delay segment are: DUs at EC, m-Wave processing, conversion, and propagation, and number of radio sub-frames.

[7]This includes the radio frequency propagation delay plus the analog component delay (which is ignored in the preliminary results of Sec. V).

[8]E.g., optical line terminal (OLT) at CC and ONU at EC

[9]In optical ring topology for cities, in each direction, the optical signal passes through a data-center optical switch, area node optical switch, and an access node optical switch. We ignore the last, i.e., access node switch, because the last segment of our proposed H-CRAN network is considered as m-Wave.

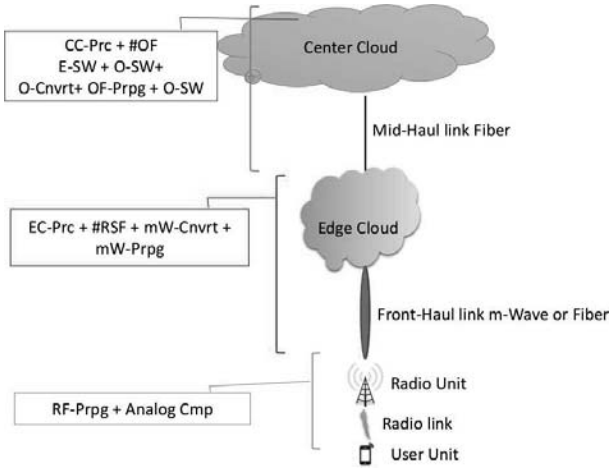

Fig. 3. Delay model for each element of the proposed architecture, given virtualized function processing (function-split).

information in Table II the amount of processing needed per communication function i per reference unit time is calculated as follows:

$$C_i = C_{i,ref} \prod_{x \in X} \left[\frac{x_{act}}{x_{ref}} \right]^{s_{i,x}}, \quad (1)$$

where x_{act} and x_{ref} are the system input parameters/resources under actual scenario and the reference scenario, respectively, and X is the set of all possible tuning parameters. The exponent $s_{i,x}$ highlights the impact of changing the input parameter on the required GOPS for the communication function sub-component. It follows that the delay induced by processing each individual function is calculated based on the equipment processing power, as below,

$$d^y_{i,prc} = \frac{C_i}{\mathcal{C}^y_{Eq}}, \quad (2)$$

where $\mathcal{C}^y_{Eq}$ is the processing power of the equipment at y, the superscript $y \in \{EC, CC\}$ is to express the equipment location at either EC or CC. The equipments' processing power at the central and edge clouds are related by an efficiency factor, $\eta_{EC} \in [0, 1]$, as follows,

$$\mathcal{C}^{EC}_{Eq} = \eta_{EC} \mathcal{C}^{CC}_{Eq}, \quad (3)$$

The delay induced by accumulating functions processing given a specific split decision, both user function split $p_{u(k)}$ (at content k) and cell function split q_c, is expressed as follows,

$$D_{prc}\left(p_{u(k)}, q_c\right) = \sum_{i \in \left[p_{u(k)}, |F_{UP}|\right]} d^{CC}_{i,prc} + \sum_{i \in \left[0, p_{u(k)}\right]} d^{EC}_{i,prc} + \sum_{i \in [q_c, |F_{CP}|]} d^{CC}_{i,prc} + \sum_{i \in [0, q_c]} d^{EC}_{i,prc} \quad (4)$$

One other major factor that contributes to the delay is the number of optical frames and radio sub-frames needed to transmit the requested content to the user through the midhaul, fronthaul, and radio link. The delay induced by the number of radio sub-frames is found as follows,

$$D_{N_{rsf}} = N_{rsf} T_{rsf} \quad (5)$$

where T_{rsf} is the is the time required to transmit one radio sub-frame, e.g., $T_{rsf} = 1$ msec. The number of required radio sub-frames, denoted as N_{rsf}, is calculated as,

$$N_{rsf} = \left\lceil \frac{V_u}{N_{SC_{sf}} N_{SYM_{sf}} u_{PRB} (1 - OH_{RP}) u_{MI}} \right\rceil \quad (6)$$

where the requested content volume is V_u, the number of sub-carrier per radio sub-frame is $N_{SC_{sf}}$, number of symbols per sub-frame is $N_{SYM_{sf}}$, the physical resource blocks allocated for the user is u_{PRB}, the overhead introduced by communication protocol is $(1 - OH_{RP})$, and the user's modulation index is u_{MI}. On the other hand, the delay induced by the number of optical frames that are needed to transport the requested content depends, at specific function split, is found as follows,

$$D_{N_{of}}\left(p_{u(k)}, q_c\right) = N_{of}\left(p_{u(k)}\right) T_{of} + N_{of}\left(q_c\right) T_{of}, \quad (7)$$

where $N_{of}\left(p_{u(k)}\right)$ is the number of optical frames needed to transport the data volume resulted from a user function processing split $p_{u(k)}$ (associated with user's content $u(k)$). T_{of} is the optical frame time. $N_{of}\left(q_c\right)$ is the number of optical frames needed to transport the volume of data resulted from a cell function processing split q_c. The calculation of $N_{of}\left(p_{u(k)}\right)$ and $N_{of}\left(q_c\right)$ depends on the transportation strategy of the central cloud to a single request after deciding to split the processing in between CC and EC. For instance, previously, we assumed that if the row data packet is transport via optical to the EC without any processing at CC, then, common fiber transportation protocol is employed and we don't care about the data resulted from different radio function processing split. However, inhere, we assume that the optical link wait for the processing of communication function, which depends on the splits option (associated by the requested content), then transport the resulted data volume of this processing. This analogy is similar to the optical burst switching concept. Based on this analogy, the number of needed optical frames to transport the data volume resulted from a user function processing split is found as follows,

$$N_{of}\left(p_{u(k)}\right) = \left\lceil \frac{V^{cc}\left(p_{u(k)}\right) N_{rsf}}{S_{of}\left(|\mathbb{C}|\right)} \right\rceil \quad (8)$$

where $V^{cc}\left(p_{u(k)}\right)$ is the data volume resulted from user's function split at CC, $S_{of}(|\mathbb{C}|)$ is the amount of bits that can be accommodated by the optical frame, divided by $|\mathbb{C}|$ number of cells. Given that several cells shares the same optical link with TWDM PON [12]. Similarly, the needed optical frames based on cell function processing split is found as follows,

$$N_{of}\left(q_c\right) = \left\lceil \frac{V^{cc}\left(q_c\right) N_{rsf}}{S_{of}\left(c\right)} \right\rceil \quad (9)$$

where $V^{cc}\left(q_c\right)$ is the data volume resulted from cell's function split at CC. Note that $V^{cc}\left(p_{u(k)}\right)$ and $V^{cc}\left(q_c\right)$ are straightforward to obtain in similar lines with [12].

Finally, In order to calculate the overall delay, which is induced by the different function split and includes all components described in Fig. 3, the following delay formulation

is proposed,

$$\begin{aligned} D_T =& D_{prc}\left(p_{u(k)}, q_c\right) + D_{N_{rsf}} + D_{N_{of}} + D_{onu} + D_{lc} \\ &+ D_{opg} + D_{mWpg} + D_{mWcnv} + D_{rpg} \\ &+ \left[\mathbb{I}\left(p_{u(k)} < |F_{UP}|\right) + 2\right] D_{sw}, \end{aligned} \tag{10}$$

where the delay of ONU, LC, optical propagation, m-Wave propagation, m-Wave conversion process, radio propagation, and switches are denoted, respectively, as: D_{onu}, D_{lc}, D_{opg}, D_{mWpg}, D_{mWcnv}, D_{rpg}, and D_{sw}. The indicator function $\mathbb{I}\left(p_{u(k)} < |F_{UP}|\right)$ is for adding one more switch delay if part of the functions processing occurred at CC.

As noticed from the aforementioned delay model components, the communication function split model, i.e., cell and user function splits, have a significant impact on the delay model. Since both cells' function processing and users' function processing impact the overall delay, the delay of the users in a cell is dominated by extreme delay requirement users. Hence, in this work, we assume that the delay of all users in a single cell is controlled by the lowest delay. Although this assumption has a logical background (in-terms of cell and user functions processing), many realistic cases can be found as use-cases for this claim. For instance, online-gaming users in an entertainment coffee-shop will require similar delay quality of service (QoS).

IV. Incorporating of Delay in Functional Split Optimization problem

In this section, we investigate the possibilities of incorporating the delay model in the minimization problems of energy and midhaul bandwidth consumption. Several incorporation options can be considered. For instance, including the delay as a constraint or including the normalized delay as another component of the objective function. Before discussing each option, we recall our previously proposed multi-objective optimization, that includes minimization of energy and bandwidth [1].

In our previous work, we have targeted an optimization problem to minimize a linearly weighted sum of the system's normalized power consumption plus normalized total bandwidth consumption of midhaul. The objective function formulation is expressed as follows,

$$\min w_p \cdot \frac{\mathcal{P}_T}{p_n} + w_b \cdot \frac{\mathcal{B}_{MH}}{b_n} \tag{11}$$

where w_p and w_b are the weighting factor of the power consumption and the midhaul bandwidth consumption, respectively. We choose w_p = 1 - w_b, i.e., to highlight the complementary impact of the associated metrics. The parameters p_n and b_n are the normalization factors of each the power and bandwidth consumptions, respectively[10]. The notations $\mathcal{P}_T$ and $\mathcal{B}_{MH}$ denotes the total power and midhaul bandwidth consumption, respectively. The total power consumption is expressed as,

$$\begin{aligned} \mathcal{P}_T =& g \cdot P_{LC} + \left(P_{CC} + l P_{CC}^{DU}\right) \mathbb{I}\left(l > 0\right) + \\ & \sum_{e \in \mathbb{E}} \left(P_{ONU} + P_{EC} \mathbb{I}\left(l_e > 0\right) + l_e P_{EC}^{DU}\right) \end{aligned} \tag{12}$$

where the power consumption of DU at CC and EC are expressed as P_{CC}^{DU} and P_{EC}^{DU}, respectively. The power consumption of LC, ONU, housing at both CC and EC are expressed respectively as P_{LC}, P_{ONU}, P_{CC} and P_{EC}. The parameters l and l_e are the number of active DU at CC and at e^{th} EC (where the integer $e \in \{0, ..., |\mathbb{E}|\}$), while g is the number of active wave-length. The midhaul bandwidth consumption is obtained by summing over all active wavelength, $w \in \{0, ..., |\mathbb{W}|\}$ induced by all ECs, i.e., $e \in \{0, ..., |\mathbb{E}|\}$ and the associated cells, $c \in \{0, ..., |\mathbb{C}|\}$, as follows,

$$b_{MH} = \sum_{w \in \mathbb{W}} \sum_{e \in \mathbb{E}} \mathbb{I}\left(w_e = w\right) \sum_{c \in \mathbb{C}_e} \left(G_c(q_c) + \sum_{u \in \mathbb{U}_c} J_u(p_{u(k)}) \right), \tag{13}$$

where $G_c(q_c)$ is a function that relates the cell processing split (q_c) split of cell c to the required midhaul bandwidth, can be found in [12]. The function $J_u(p_u)$ relates the user processing split, p_u (of the u's user, where $u \in \{0, ..., |\mathbb{U}_c|\}$), to the required midhaul bandwidth, can be found in [12].

Following the aforementioned function splits model and its impact on the system's energy and bandwidth consumption, we now investigate the possibilities and impacts of incorporating the user's request delay in the function split model. Different approaches could be considered as follows.

1) Variety of delay requirement per user's service should be considered. For instance, to satisfy the user's requirement, the actual end-to-end delay per request, which resulted from the split decision, must meet this delay requirement by enforcing an additional constraint to the original problem. This constraint might result in enforcing different function splits (at UP or CP) to meet the user's delay requirement. Different split decision might result in activating more/less DUs at either EC or CC. Hence, it directly impacts the energy and midhaul bandwidth consumption. In Sec. V we include a numerical figure, which describes how different split decisions impact the system delay performance.
2) Incorporating the actual end-to-end delay per user request in the objective function to be minimized in combined with the normalized energy and midhaul bandwidth, as in [1]. As described in the previous point, the optimal function processing split point that minimizes the system energy might not be the optimal point for minimizing the user's request delay performance. Hence, a trade-off among system's energy, midhaul's bandwidth, and delay can be deeply investigated by including the normalized delay in the objective function.
3) Employing a cross-segment optimization, which considers the delay requirement in jointly allocating and scheduling the available resources at both transport and radio access end. For instance, when employing delay related techniques, e.g., soft hybrid automatic repeat

[10]The normalization factors, p_n and b_n are the maximum consumed power and maximum consumed midhaul bandwidth, respectively.

request (H-ARQ) decision, users with similar delay requirements should be allocated to share the same virtual cell equipments.

V. Study Case

In this section, we describe a use case for the impact of communication function processing split options on the individual component delay and overall delay. The use case considers a single user scenario under similar configuration to the Long-Term Evolution (LTE) frame structure. Note that since both cells' and users' function split do not overlap, we map both of them into a single function split that count from 0, which corresponds to CPS-1, up to 8, which corresponds to UPS-n+1, based on the notation of Fig. 2.

TABLE I
Simulation Parameters.

Parameter Name	Value
Radio Sub-Frame (RSF) Time	1e-3
Number of Symbols in RSF	12
Number of Sub-Carriers in RSF	14
Transport block size	57376
Bandwidth	20 MHz
MIMO	2x2
Delay of Optical transmission[11]	$\approx$ 0.4 * 1e-3
Ethernet switching delay	$<$ 5.2 * 1e-6 [16]
m-Wave conversion delay	30.*1e-6
Optical switching delay	$\approx$ 2.5 * 1e-3 [17]
Reference computation power of CC's Equipment	100 GOPS
$\|\mathbb{C}\|$	4
$S_{of}(\|\mathbb{C}\|)$	38880 * 8 / $\|\mathbb{C}\|$
T_{of}	125 * 1e-6
User's request size	1 Mbit

Figure 4 and Fig. 5 evaluate the delay induced from several components of the system model, which have been described earlier in Fig. 3, versus all function split options. Fig. 4 evaluates the delays for an efficiency of $\eta_{ec} = 1.0$, whereas, Fig. 5 evaluates the delays for an efficiency of $\eta_{ec} = 0.8$. In the x-axis, going to the left ($0 \leftarrow$) infers that more processing functions are centralized at CC, whereas, going to the right ($\rightarrow 8$) infers that more processing functions are send toward EC. At Fig. 4, it is expected to find that the total delay decreases with higher split option, because the computation capabilities of both EC and CC devices are the same. Whereas, in Fig. 5, the delay increase then decreases with the increase of split value, due to the low efficiency in processing power of EC equipment in compared to CC ones. Intuitively, the delay induced by EC's function processing component increases with the split option, whereas the delay induced by CC's processing decreases with higher split option. It is also interesting to note that the delay induced by the number of optical frames (N_{OF}) needed to transport the data volume at the associated split decreases with the increment of the split. On the contrary, the delay induced by the number of radio sub-frames (N_{RSF}) is constant with respect to different split options since it is not a function of it, as explained in the Sec. (III).

Figure 6 evaluates the total delay performance versus all function processing split options, for variety of resource blocks

[11]This delay is a result of optical transmission related parameters in OLT and ONU devices, e.g., processing, power amplifier, conversion between electrical to optical signals, and coding for optical transmission [15].

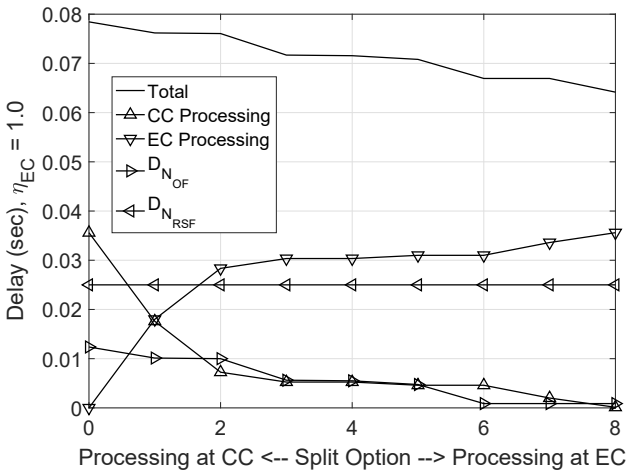

Fig. 4. Delay performance of different contributing components versus all function processing split options, at $\eta_{EC} = 1$.

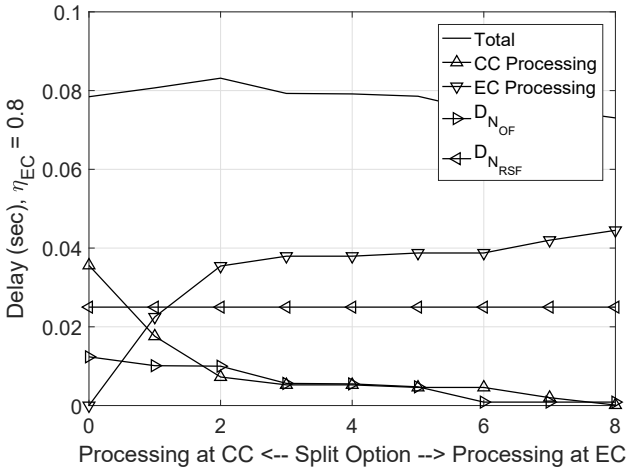

Fig. 5. Delay performance of different contributing components versus all function processing split options, at $\eta_{EC} = 0.8$.

and equipment computation efficiency parameters. For low η_{EC} the total delay increases then decreases with higher split option because processing at EC might save transportation delay but have higher processing delay. Whereas, at high η_{EC} the delay decreases with higher split as expected. Intuitively, higher allocation of resource blocks to a user, i.e., u_{PRB}, induces lower delay.

VI. Conclusion

In this work, we proposed an end-to-end delay model, per user's request, that quantifies the impact of function split options on the user's request delay performance. The model is based on our previously proposed architecture of a hybrid CRAN with different base band function splits. In this delay model, we take into account all the delay components in all network's segments, i.e., central cloud, edge cloud, radio unit, midhaul link, fronthaul link, and radio link. It is noted that meeting the user's delay requirements leads to different function split options, hence (de)activating digital units at either central or edge cloud, which directly affects the energy consumption of the system. Therefore, we proposed potential

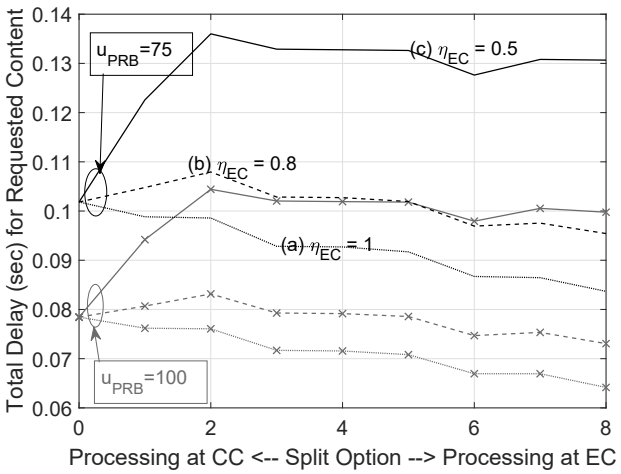

Fig. 6. Total delay performance versus all function processing split options, under variable $\eta_{EC} = \{0.5, 0.8, 1\}$, $u_{PRB} = \{75, 100\}$, and $N_{SYM_{sf}} = \{14\}$.

directions to incorporate this delay model in the existing frameworks of minimizing the energy and midhaul's bandwidth consumptions and study their trade-off. We investigated the overall delay and the breakdown of each component contribution. It is found that the behavior of the total delay is highly impacted by the amount of processing at the edge cloud and the efficiency ratio between the edge and central cloud equipment's processing power. At high processing efficiency ratio (≈ 1) processing more functions at the edge cloud will always improve the delay.

Appendix A
Reference Table for Complexity and Exponents of Communication Functions

Table II describes the amount of GOPSs required for each digital component of the communication functions processes. It also provides the scaling factor exponents which highlight the impact of change different parameter on the original reference case, i.e., SISO, 20 MHz, 6 bps/Hz (64-QAM, coding rate 1).

TABLE II
Downlink reference complexity and exponent of digital components for SISO, 20 MHz, 6 bps/Hz (64-QAM, coding rate 1)

Subcomponent	**GOPS**	**Scaling Exponents** ($s_{i,x}$)				
		BW	**Mod Ind**	**Ant.**	**Load**	**Str**
CPRI	18	1	1	1	1	1
Predistortion	10.7	1	0	1	0	0
Filtering	6.7	1	0	1	0	0
Up/Down-sampling	2	1	0	1	0	0
TD non-ideal. est./comp.	1.3	1	0	1	0	0
FFT/IFFT, FD non-ideal.	4	1.2	0	1	0	0
Synchronization	0	0	0	1	0	0
Channel estimation	0	1	0	1	0.5	1
Equalizer Comp.	0	1	0	3	1	0
Equalization	0	1	0	2	1	0
Mapping/Demapping	1.3	1	1.5	0	1	1
OFDM Mod./Demod.	1.3	1	0	1	0.5	0
MIMO precoding	1.3	1	0	1	1	1
Channel coding	5.2	1	1	0	1	1
Control	2.7	0	0	0.5	0	0.2
Upper Network Layer	8	1	1	0	1	0

References

[1] X. Wang, A. Alabbasi, and C. Cavdar, "Interplay of energy and bandwidth consumption in cran with optimal function split," in *To appear in proceedings of IEEE International Conference on Communication (ICC)*, 2017.

[2] C. L. I, J. Huang, R. Duan, C. Cui, J. . Jiang, and L. Li, "Recent progress on c-ran centralization and cloudification," *IEEE Access*, vol. 2, pp. 1030–1039, 2014.

[3] A. Checko, A. C. Juul, H. L. Christiansen, and M. S. Berger, "Synchronization challenges in packet-based cloud-ran fronthaul for mobile networks," in *2015 IEEE International Conference on Communication Workshop (ICCW)*, June 2015, pp. 2721–2726.

[4] S. Khalili and O. Simeone, "Uplink harq for cloud ran via separation of control and data planes," *IEEE Transactions on Vehicular Technology*, vol. PP, no. 99, pp. 1–1, 2016.

[5] J. Li, M. Peng, A. Cheng, Y. Yu, and C. Wang, "Resource allocation optimization for delay-sensitive traffic in fronthaul constrained cloud radio access networks," *IEEE Systems Journal*, vol. PP, no. 99, pp. 1–12, 2014.

[6] A. Checko, A. P. Avramova, M. S. Berger, and H. L. Christiansen, "Evaluating c-ran fronthaul functional splits in terms of network level energy and cost savings," *Journal of Communications and Networks*, vol. 18, no. 2, pp. 162–172, April 2016.

[7] M. Artuso, A. Marcano, and H. Christiansen, "Cloudification of mmwave-based and packet-based fronthaul for future heterogeneous mobile networks," *IEEE Wireless Communications*, vol. 22, no. 5, pp. 76–82, October 2015.

[8] M. A. Khalighi and M. Uysal, "Survey on Free Space Optical Communication: A Communication Theory Perspective," *IEEE Communications Surveys Tutorials*, vol. 16, no. 4, pp. 2231–2258, Fourthquarter 2014.

[9] "40-Gigabit-capable passive optical networks (NG-PON2)," ITU-T G.989 series of Recommendations:, ITU-T, March 2013.

[10] J. Liu, S. Zhou, J. Gong, Z. Niu, and S. Xu, "Statistical multiplexing gain analysis of heterogeneous virtual base station pools in cloud radio access networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 8, pp. 5681–5694, Aug 2016.

[11] T. Pfeiffer, "Next generation mobile fronthaul and midhaul architectures [invited]," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 7, no. 11, pp. B38–B45, November 2015.

[12] "Functional splits and use cases for small cell virtualization." Release, Small Cell Forum, Jan. 2016.

[13] "IEEE P1914.1 Meeting Materials. [online]:http://sites.ieee.org/sagroups-1914/," IEEE P1914.1 TF meeting materials, IEEE, August, 2016.

[14] B. Debaillie, C. Desset, and F. Louagie, "A flexible and future-proof power model for cellular base stations," in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, May 2015, pp. 1–7.

[15] F. Aurzada, M. Scheutzow, M. Reisslein, N. Ghazisaidi, and M. Maier, "Capacity and delay analysis of next-generation passive optical networks (ng-pons)," *IEEE Transactions on Communications*, vol. 59, no. 5, pp. 1378–1388, May 2011.

[16] Siemens. (2017) Website. [Online]. Available: https://w3.siemens.com/mcms/industrial-communication/en/rugged-communication/Documents/AN8.pdf1

[17] ARRIS. (2017, Jun.) Website. [Online]. Available: http://www.arris.com/globalassets/resources/data-sheets/ch3000-os3200-series-optical-switches-data-sheet.pdf