Delay Constrained Hybrid CRAN: A Functional Split Optimization Framework

*Abdulrahman Alabbasi, *Miguel Berg, and †Cicek Cavdar
Email: *{abdulrahman.alabbasi,miguel.berg}@ericsson.com, †{cavdar}@kth.se

***Abstract*—Hybrid cloud radio access network (CRAN) architecture supports the realization of splitting/virtualization of the baseband processing unit (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 processing efficiency. In our previous work, we studied the impact of different split points (which decide the amount of communication processing at center and edge clouds) on the system's energy and midhaul link bandwidth consumptions. Midhaul is defined as the part of the network connecting center cloud to edge cloud. In this study, we propose an optimization framework to incorporate the end-to-end delay, from the central cloud to the end user, under different/flexible function split points. For all services requests, all users have the same services' size (in Mbits) that are associated with different delay requirements based on the service. Different service/delay requirements enforce different function splits. Our proposed optimization framework minimizes both system's power and bandwidth consumption with guaranteed end-to-end latency performance (control plane is not considered). The required split decision is significantly dependent on the processing power efficiency ratio between processing units at edge and central clouds. As shown in our previous results, Hybrid CRAN achieves about 55 percentage power saving compared fully distributed system, at the expense of midhaul bandwidth consumption. Simulation results showed that as the delay requirements decrease, average allowable latency reduces from 65msec to 45msec, this power consumption (achieved by relaxed delay case) increases by average of 33 percentage, at 20 resource blocks per user.**

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

I. Introduction

Radio access networks will evolve towards denser deployments to meet with the high capacity requirements of the next generation mobile services. CRAN architecture is initially proposed as a cost efficient and scalable solution to centralize the base band unit (BBU) processing to reduce the capital and operational expenses (CAPEX and OPEX) due to densification [1]. CRAN enables sharing of processing resources thanks to virtualization of functions and provides a multiplexing gain for saving system's energy consumption and cost. However it imposes high bandwidth and stringent delay requirement on the fronthaul link, the part of the network connecting centralized cloud to the radio units. To relax these requirements, the concept of communication function splitting is proposed to allow partial centralization of network functions [2]. The splitting concept enables the processing of communication functions at multiple sites. For instance, a split point in the communication stack, is a point where the communication functions after this point are processed closer to the user, and the functions before this point are processed in a central node (far from the users). In this work, we investigate the impact of delay on the system performance and on the optimal allocation of communication functions processing at centralized site or distributed sites. For this purpose, we integrate an end-to-end delay model into the constraint programming optimization framework. This delay is defined as the time needed to deliver a certain content requested by the user from Central-Cloud (CC) to the user's terminal. This includes the delay induced by processing at CC and/or Edge-Cloud (EC), midhaul and fronthaul transportation delay, and radio access transmission delay.

Several works have studied the delay performance of CRAN in addition to proposal of protocols to realize a flixible functional split. Authors of [3] have proposed an architecture framework and ethernet-based protocol to support flexible function split. Their protocol supports different hybrid CRAN topologies and split options, in addition to several key performance metrics. Early results on the field trials of CRAN's delay (with(out) the Wavelength Division Multiplexing (WDM) optical ring) are reported in [4]. The authors of [5] 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. Authors of [6] 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 radio unit (RU). In CRAN settings, authors of [7] have proposed a queue-aware power and rate allocation for delay-sensitive traffic and formulate it as a Markov decision process. Authors of [8] evaluated the impact of certain function splits, i.e., Physical, media access control (MAC), and packet data convergence protocol (PDCP)-radio link control (RLC) splits, on the energy and cost savings of the CRAN network. The Teletraffic theory has been used in the aforementioned quantitative study. This study considered a two-sites processing architecture with the remote sites being the basestations. Unlike our work, none of the above literatures has considered the impact of delay performance of end users' requests on the variable function processing splits between edge cloud and centralized cloud. There is no study on the end-

978-1-5386-4920-6/18/$31.00 ©2018 IEEE

to-end delay considering all the optical and wireless segments of the network from the central cloud to the user except [9], where we have published preliminary results. In our study, end-to-end delay is defined as the delay needed to deliver a certain content requested by the user from CC to the user's terminal. This includes the delay induced by processing at CC and/or EC, midhaul and fronthaul transport delay, and radio access transmission delay. Transport delay includes conversion from digital to optical, vice versa, and from digital to mmWave (milli-meter wave).

In this work, we consider the architecture of hybrid cloud radio access network (H-CRAN), proposed in [10]. In H-CRAN architecture, digital units (DUs) are deployed at both ECs and CC to allow processing of mobile network functions in each or both of CC and/or ECs. Our objective is to find the optimal function split point that minimizes a weighted sum of system's power and midhaul bandwidth consumptions. Given the assumption that centralizing processing saves power[8] [1], whereas, distributing processing saves midhaul bandwidth. We also optimally allocate the DUs and wavelengths, which minimize the objective, while meeting the DU processing and midhaul link capacities. Resources are optimally allocated in all segments to satisfy the heterogeneous users' service requirements translated into end-to-end delay. Also, this delay model takes into account the associated delay of different function processing split points. It considers all the delay components induced by the network's segments from CC to EC to RU, then to user equipment (UE). For each user the delay induced by the selected function processing split point is constrained by the targeted user's request delay.

Thanks to the integration of end-to-end delay constraint into the constrained programming model and solving the model optimally, we can decide network function splits per users' service, which is called service-oriented functional split. Also, our framework enables the analysis of interplay between delay, energy, processing resources and bandwidth consumption. It is shown that hybrid CRAN achieves a maximum of 55 % power saving compared to fully distributed system or a maximum of 77 % midhaul bandwidth saving compared to fully centralized system [10]. As services requirements vary among users, we formulate two random sets of delay requirements. The strict delay set varies within [29,60] msec, whereas relaxed delay set varies within [29,100] msec. Satisfying services that fall under the strict delay set (i.e., more users requests low latency services) resulted in increasing power consumption (in compared to relaxed delay requirements) by an average of 32.8%.

II. Network Architecture

In here, we present the system architecture and the function processing splits options, in line with our work in [10].

We consider H-CRAN architecture, i.e., 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 is implemented using milli-meter Wave links [11][2]. The ECs are connected to CC via midhaul using time and wavelength division multiplexing (TWDM)-passive optical network (PON) [12], 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 CC and EC[3]. 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 services and network functions. Hence, the DUs are capable of sharing their computational resources among several connected RUs (if implemented in general purpose servers). However, EC is usually less energy- and computationally- efficient than CC, because larger number of equipments can share the same cooling device at CC. Also, higher number of DUs can be linked to the RUs, while at at the EC, since the space is limited, limited number of DUs can be allocated. Additionally, CC's equipments are easier to be updated, hence expected to have higher computational power. 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 (to save energy or improve delay, depending on the computational efficiency).

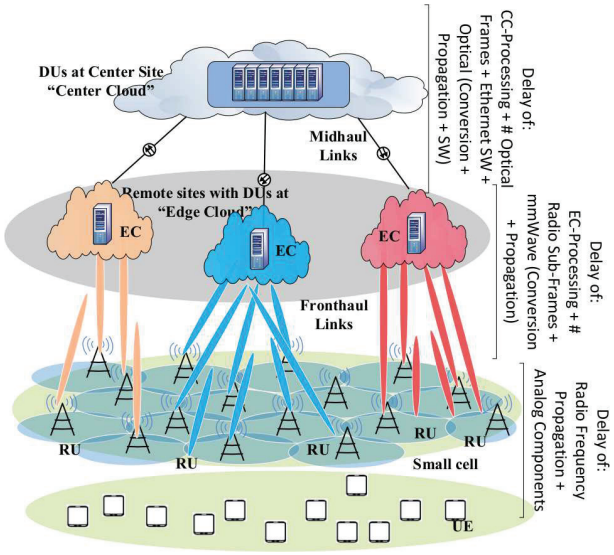

Fig. 1. H-CRAN architecture [10], and details of the considered delay model's components.

The approach we use for function processing distribution or

[1]This assumption is valid for several reasons. First, larger number of users sharing an upgraded digital unit for communication function processing, leads to higher processing energy efficiency. Second, sharing infrastructure (especially cooling) energy consumption among all users is more efficient than allocating individual units per cell or edge cloud.

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

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

centralization is similar to that in our work [10]. The 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[4]. CPs are a sequence of functions in physical layer that are dedicated for processing signals from a cell, when signals of UEs are multiplexed The per-cell processing will be terminated at CP_m. Then, UPs is a sequence of functions that will continue to process the signal streams on a per-UE basis including (1) equalization, Fourier transform, (2) modulation/demodulation, MIMO (de)mapping and (pre)coding, (3) forward error correction, turbo decoding, and (4) upper layers functions. Functional split can occur before CP_1, after UP_n, or in-between any two functions. Note that CP split 1 (CPS_1) expresses the fully centralized CRAN. Whereas, UP split n+1 (UPS_{n+1}) expresses a fully distributed baseband processing deployment, i.e., current cellular systems.

In the proposed system model, we enable several functionalities that improve both power and midhaul bandwidth savings.

- We enable DUs shut down when users' processing load at a site is low enough.
- We enable shutting down the cooling equipment at ECs or CC given that no DU is active at that site.
- We enable shutting down midhaul, fronthaul, and radio access components at low load where the EC do not have any active user.
- We enable shutting down a wavelength, given that the associated cells do not have any active user.

III. Delay Model Based on Function Split

In this section, we present our proposed end-to-end delay model per user's request. Figure 1 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 allocation of the communication functions processing. Different function splits result in different amount of processing at CC and EC, 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 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 and/or EC. This processing is conducted based on each radio sub-frame, and 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. Each split results in a larger amount of data (than the requested one) which needs N_{of} optical frames to be transported via midhaul link.
- Accumulative delay induced by encapsulating the requested content over multiple radio sub-frames, and pushed to m-Wave link (fronthaul) which then will be pushed to radio frequency-link to the end user.
- Constant delay induced by optical and m-Wave conversion and propagation.
- Constant delay induced by optical switch at the data-center and the area nodes, 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., $N_{of} \approx 1$.

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). Also, the user processing delay is not considered in this work, due to the lack of measurements about the computational capability of the hand held devices.

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. In our work [9], we listed all the digital sub-components, which contribute to the overall delay, the associated GOPS per each communication sub-component function, and the associated exponent factors (that shows the impact of using radio parameters that differ from the reference, i.e, modulation index, resource blocks, etc.). Utilizing these information, the amount of processing needed per communication function i per reference unit time is calculated as follows,

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

where x_{act} and x_{ref} are the system input parameters/resources under actual and 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 communication function 'i'. It follows that the delay induced by processing each individual function is calculated based on the equipment processing power, as below,

$$d^{y}_{i,prc} = \frac{C_i}{\mathcal{C}^{y}_{Eq}}, \tag{2}$$

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 equipment's processing power at CC and ECs are mathematically related by an efficiency factor, $\eta_{EC} \in [0, 1]$, as follows,

$$\mathcal{C}^{EC}_{Eq} = \eta_{EC} \mathcal{C}^{CC}_{Eq}, \tag{3}$$

The delay induced by processing the communication func-

[4]In this study, $m = 4$ and $n = 4$, as described in [2], [13], [14].

[5]This includes 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]This includes DUs at EC, Milli-Meter wave (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.

tions, given a specific split decision and a content k of user u, is expressed as follows,

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

where $p_{u(k)}$ is a user function processing split and q_c is a cell function processing split. Note that both $p_{u(k)} \in [0, |F_{UP}|]$ and $q_c \in [0, |F_{CP}|]$ are rigorously defined in Sec. IV-B.

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,

$$D_{N_{rsf}} = N_{rsf} T_{rsf} \tag{5}$$

where T_{rsf} 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,

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

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,

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

where $N_{of}\left(p_{u(k)}\right)$ is the number of optical frames needed to transport the data volume resulted from split $p_{u(k)}$. 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 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 [8]. The number of optical frames needed to transport the data of a specific function split is obtained as follows,

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

where $V^{cc}\left(p_{u(k)}\right)$ is the data volume resulted from cell or user function split at CC of each radio sub-frame, $S_{of}(|\mathbb{C}_e|)$ is the amount of bits that can be accommodated by the optical frame, divided by $|\mathbb{C}_e|$ number of cells in EC e[9].

In order to calculate the overall delay, which is induced by the different function split and includes all components described in Fig. 1, the following delay formulation is considered,

$$\begin{aligned} D_T(u,c) =& D_{prc}\left(p_{u(k)}, q_c\right) + D_{N_{of}}\left(p_{u(k)}, q_c\right) + D_{N_{rsf}} \\ &+ D_{onu} + D_{lc} + D_{opg} + D_{mWpg} + D_{mWcnv} \\ &+ D_{rpg} + \left[\mathbb{I}\left(p_{u(k)} < |F_{UP}|\right) + 2\right] D_{sw}, \end{aligned} \tag{9}$$

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.

IV. Delay Constrained Optimization Problem

In this section, we aim to jointly minimize the system's power and midhaul's bandwidth consumptions, while meeting user's quality of service (QoS). This optimization problem satisfies several constrains related to the network limits, but most importantly it satisfies users' delay requirements. The aforementioned constrained optimization problem is formulated as a constraint programming [15].

A. Given

- Topology: CC is connected to many ECs. Each EC is connected to a set of cells, and a cell covers a set of UEs.
- $\mathbb{U}_x$: set of UEs. When $x = 0$, it means all UEs in H-CRAN, otherwise it refers to a set of UEs in cell $x = c$.
- $\mathbb{C}_x$: set of cells (RUs). When $x = 0$, it refers to all cells in H-CRAN, $x = e$ refers to set of cells belonging to EC e.
- $\mathbb{D}_x$: set of DUs. When $x = 0$, it refers to all DUs in H-CRAN, $x = -1$ refers to set of DUs in the CC , $x = e$ refers to set of DUs in EC e.
- $\mathbb{E}$: set of all ECs.
- $\mathbb{W}$: set of wavelengths.
- K: bandwidth capacity of a wavelength. Note that this is different from the bandwidth induced and consumed by user's and cell's function split.
- D^u_{thr}: the required delay threshold for user u.
- F_x: set of function split options, x is UP or CP splits.
- L^y_x: capacity of a DU located at y site, in terms of the number of $x = CP$ and $x = UP$ functions that can be accommodated by this DU.
- We also define an indicator function, $\mathbb{I}(.)$, as follows,

$$\mathbb{I}(a) = \begin{cases} 1; & \text{if event a is correct,} \\ 0; & \text{if event a is not correct.} \end{cases} \tag{10}$$

B. Integer Variables

- $p_u \in [0, |F_{UP}|]$: UP split of UE u. If UP of UE u is not split, then $p_u = |F_{UP}|$, otherwise, $p_u \in [0, |F_{UP}|)$.
- $q_c \in [0, |F_{CP}|]$: CP split of cell c. If CP of cell c is not split, then $q_c = |F_{CP}|$, otherwise, $q_c \in [0, |F_{CP}|)$.
- $m_u \in D_e$: DU hosting UPs of UE u at EC e.
- $n_u \in D_{CC}$: DU hosting UPs of UE u at CC.

[8] For more details please check [9]

[9] Given that several cells share the optical link with TWDM-PON [2]

- $x_c \in D_e$: DU hosting CPs of cell c at EC e.
- $y_c \in D_{CC}$: DU hosting CPs of cell c at CC.
- w_e: wavelength used by EC e.

C. Objective

Our target is to minimize a weighted sum of the system's normalized power and midhaul bandwidth consumption as,

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

where w_p and w_b are weighting factors 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. Parameters p_n and b_n are the normalization factors of each the power and bandwidth consumptions, respectively[10]. Total power and midhaul bandwidth consumptions are denoted as $\mathcal{P}_T$ and $\mathcal{B}_{MH}$, respectively. The total power consumption is expressed as,

$$\begin{aligned}\mathcal{P}_T = &\left(P_{CC} + lP_{CC}^{DU}\right)\mathbb{I}\left(l>0\right) + gP_{lc} + \\ &\sum_{e\in\mathbb{E}}\Bigg[\left(\sum_{c\in\mathbb{C}_r}\left(P_{Tx}+P_{FH}\right)\mathbb{I}(|\mathbb{U}_c|>0)\right) \\ &+\left(\mathbb{I}\left(\sum_{c\in\mathbb{C}_e}|\mathbb{U}_c|>0\right)P_{onu} + P_{EC}\mathbb{I}\left(l_e>0\right) + l_e P_{EC}^{DU}\right)\Bigg]\end{aligned} \tag{12}$$

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, fronthaul and radio links transmissions, housing at both CC and EC are expressed respectively as P_{lc}, P_{onu}, P_{FH}, P_{Tx}, P_{CC}, and P_{EC}. The parameters l and l_e are the number of active DUs at EC and e^{th} EC (where the integer $e \in \{0, ..., |\mathbb{E}|\}$), while g is the number of active wavelengths. The midhaul bandwidth consumption is obtained by summing over all active wavelengths, $w \in \{0, ..., |\mathbb{W}|\}$ induced by all ECs and the associated cells, $c \in \{0, ..., |\mathbb{C}_e|\}$, as follows,

$$\begin{aligned}\mathcal{B}_{MH} = \sum_{w\in\mathbb{W},e\in\mathbb{E}}\mathbb{I}\left(w_e=w\right)\sum_{c\in\mathbb{C}_e}\Bigg(&G_c(q_c)\mathbb{I}(|\mathbb{U}_c|>0) \\ &+\sum_{u\in\mathbb{U}_c}J_u(p_u)\Bigg),\end{aligned} \tag{13}$$

where $G_c(q_c)$ is a function that relates q_c to the required midhaul bandwidth [2]. The function $J_u(p_u)$ relates the user processing split, p_u (of the u's user), to the required midhaul bandwidth, can be found in [2]. The term $\mathbb{I}\left(w_e = w\right)$ ensures that the current wavelength belong to the remote site e.

The calculation of overall power, $\mathcal{P}_T$ in (12), clearly describes the functionalities of power saving that has been mentioned in the previous section. The terms $\left(P_{CC} + lP_{CC}^{DU}\right)\mathbb{I}\left(l>0\right)$ and $P_{EC}\mathbb{I}\left(l_e>0\right) + l_e P_{EC}^{DU}$ represent shutting down the site if there is no active DU in it. Whereas, terms lP_{CC}^{DU} and $l_r P_{EC}^{DU}$ represent shutting down the inactive DUs. The terms $\mathbb{I}\left(\sum_{c\in\mathbb{C}_e}|\mathbb{U}_c|>0\right)P_{onu}$ and gP_{lc} represent shutting down the LC and ONU if there is no active users in the associated EC. Finally, the term $\sum_{c\in\mathbb{C}_e}(P_{Tx}+P_{FH})\mathbb{I}(|\mathbb{U}_c|>0)$ represents shutting down the fronthaul and radio access transmission components if there is now active user in the associated cell.

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

D. Constraints

In this sub-section, we explain the constraints of the problem, which are divided into three types based on their functionality. First constraint captures the delay impact on the decision of function split. Second set of constraints is about satisfying the capacity limits of wavelength link's bandwidth and processing capacity of DUs. Third set of constraints ([10], omitted here due to page limit) is about guaranteeing to have a one occurrence of function split at a time.

The constraint which satisfy the user's delay requirement by allocating different function split is expressed as follows,

$$D_T(u,c) \leq D_{thr}^u\ , \forall u \in \mathbb{U}_c,\, c \in \mathbb{C}_0 \tag{14}$$

$D_T(u,c)$ is described in detail in (9). Constraint (14) is active, if the desired delay, D_{thr}^u, is within the delay range induced by different options of the function split. It is inactive constraint, if D_{thr}^u is larger than the delay range induced by function split options. It is infeasible constraint, if D_{thr}^u is lower than the delay induced by processing location.

The DUs processing capacity constraints are expressed as,

$$\sum_{e\in\mathbb{E}} H_y^x(p_u,q_c)\mathbb{I}(z_e=d) \leq L_x^y,\ \forall d \in D_0,\ \forall x \in \{UP, CP\}, \tag{15}$$

where z is the targeted DU, which is expressed as $z \in \{m_u, n_u, x_c, y_c\}$ as defined earlier in Sec. IV-B. Constraint (15) ensures that the number of UPs/CPs that are accommodated by a DU d cannot exceed DU's processing capacity L.

The wavelength capacity constraint is expressed as follows,

$$\sum_{e\in\mathbb{E}}\mathbb{I}\left(w_e=w\right)\sum_{c\in C_e}\left(G_c(q_c)+\sum_{u\in\mathbb{U}_c}J_u(p_u)\right) \leq K, \forall w \in \mathbb{W} \tag{16}$$

This ensures that the total occupied bandwidth in a wavelength cannot exceed the wavelength's capacity. The occupied bandwidth in wavelength w is the sum of the bandwidth consumptions of all ECs that are using w. The bandwidth consumption of EC e is the sum of bandwidth consumptions of all cells belong to it ($c \in C_e$). The bandwidth consumption of cell c composes of that incurred by both CP and UP splits, i.e., $G_c(q_c)$ and $J_u(p_u)$.

V. Numerical Results

In this section, we evaluate the system performance via several metrics, i.e., power, midhaul bandwidth and achievable end-to-end delay. The evaluation parameters are shown in Table I. Users are uniformly distributed in the whole area. It is assumed that user's delay threshold, D_{thr}^u, follows a uniformly distributed random variable, and its independent among users. The power and bandwidth performances are studied under

two random sets of delay thresholds, i.e., strict and relaxed sets, which are defined later in this section. We solve the proposed formulation, described in Sec. IV, via the constraint programming tool in IBM ILOG CP solver. We run the solver until the optimal values are found.

TABLE I
SIMULATION PARAMETERS.

Parameter Name	Value
Topology:4-level hexagonal	Single CC, $\|\mathbb{E}\|$=6, $\|\mathbb{C}\|$ = 43, $\|\mathbb{U}_c\| = 4$.
Configuration of RU	20 MHz, 2x2 MIMO, 64 QAM (=u_{MI}).
Capacity of DU at EC	4 CPs and 16 UPs.
Capacity of DU at CC	36 CPs and 144 UPs.
P_{EC}^{DU}, P_{CC}^{DU}, P_{EC}, P_{CC}	50, 100, 250, 500 W
P_{lc}, P_{onu}, $P_{Tx} + P_{FH}$	20, 5, 20 W
T_{rsf}, T_{of}	1e-3, 125e-6 sec
$N_{SYM_{sf}}$, $N_{SC_{sf}}$, u_{RRB}, OH_{RP}	12, 14, 20, 0.08
Delay of Optical transmission[11]	$\approx$ 0.4e-3 sec
Ethernet & Optical switching delay	$\approx$ 5.2e-6 sec[17]
m-Wave conversion delay	30e-6 sec
$\mathcal{C}_{Eq}^{CC}$	300 GOPS
$S_{of}(\|\mathbb{C}_e\|)$	38880 * 8 bits
Content's size	1 Mbit

Figure 2 evaluates the delay induced from several components of the system model versus all function split options, at an efficiency of $\eta_{ec} = 1.0$. Low value of split option ($\rightarrow 0$) refers to more processing in CC, higher value refers to more processing in EC. As expected, in Fig. 2, we find that the total delay decreases with higher split option, because the computation capabilities of both EC and CC devices are the same. Intuitively, the delay induced by EC's function processing increases with the split option, whereas the delay induced by CC's processing decreases with higher split option. It is noted that the delay induced by N_{of} decreases with the increment of the split option. On the contrary, the delay induced by N_{rsf} is constant with respect to different split options.

Figure 3 evaluates the total delay performance versus all function processing split options, for different resource block allocated per user, $u_{PRB} \in \{20, 50\}$, and different equipment computation efficiency, $\eta_{EC} \in \{0.5, 0.8, 1.0\}$. At 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 $\eta_{EC} = 1$, the delay decreases with higher split, because processing capability at both EC and CC is the same. Intuitively, higher allocation of resource blocks to a user, i.e., u_{PRB}, induces lower end-to-end latency.

Figure 4 evaluates the normalized system performance metrics, i.e., objective function ($\text{Obj}_{\text{S,R}}$), power consumption ($\text{Power}_{\text{S,R}}$), required midhaul bandwidth ($\text{Band}_{\text{S,R}}$), and the amount of function processing at EC ($\text{FP@EC}_{\text{S,R}}$), versus different power weight values, w_p. In general, higher w_p values puts more weight on the minimization of power, hence the optimal split point centralizes more processing at CC. This

[11]This delay is a result of optical transmission related parameters in optical line terminal (OLT) and ONU devices, e.g., processing, power amplifier, conversion between electrical to optical signals, and coding for optical transmission [16].

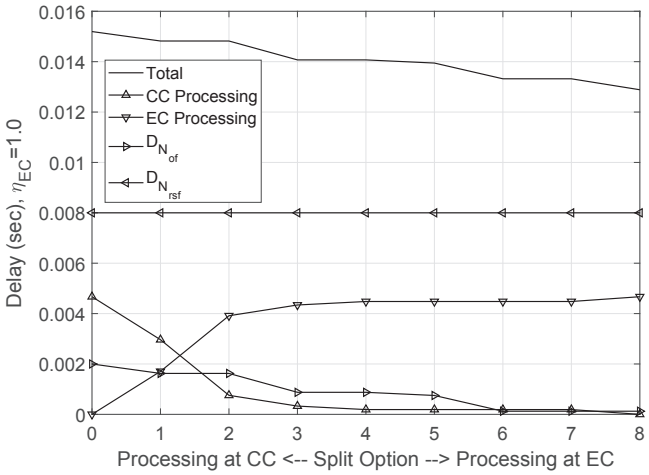

Fig. 2. Delay breakdown of all components versus function split options, at $\eta_{EC} = 1$.

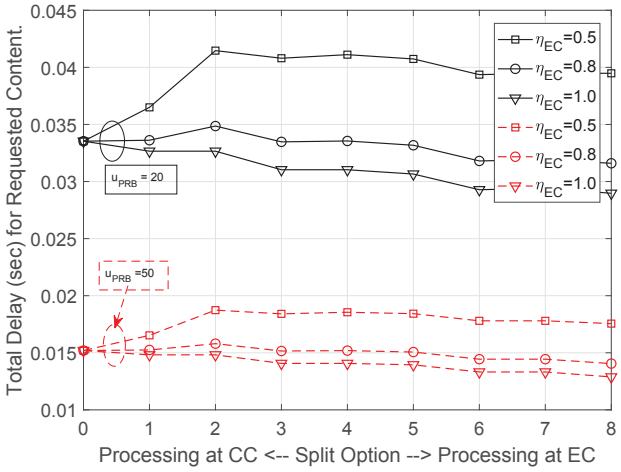

Fig. 3. Total delay performance versus function splits, under variable η_{EC} and u_{PRB}.

is observed by the decreasing behavior of $\text{FP@EC}_{S,R}$, which expresses the percentage of processing functions at EC (it is complementary to percentage of processing functions at CC, i.e. $\text{FP@CC}_{S,R} = 1 - \text{FP@EC}_{S,R}$). The {S,R} subscripts of $\text{Obj}_{\text{S,R}}$, $\text{Power}_{\text{S,R}}$, $\text{Band}_{\text{S,R}}$, and $\text{FP@EC}_{\text{S,R}}$, correspond to strict and relaxed sets of delay requirements. In Fig. 4, we consider two sets of uniformly distributed delay thresholds per users, D_{thr}^{u}, i.e., the values of each set are drawn from a uniformly distributed random variable[12]. The first set of delay thresholds, noted in the figure as 'S' (Strict) subscript, is randomly selected as $\{S : D_{thr}^{u} \in [0.029, 0.06]\}$, whereas the second set, noted as 'R' (Relaxed) subscript, is randomly selected as $\{R : D_{thr}^{u} \in [0.029, 0.1]\}$. The minimum value of the delay threshold range, i.e., 0.029 in this setting, is selected to meet the minimum possible delay that can be achieved

[12]This random model is to capture wide spectrum of delay requirements.

by the best split option for this user's available resources and requested service. Whereas, the upper bound value of the delay range, i.e., 0.06 or 0.1 in this setting, is chosen to increase the variance of the selected delay thresholds. Serving users with stricter delay requirements (average value of 44.5 msec) has significant impact on the system's power and midhaul bandwidth consumptions, compared to relaxed delay requirement case (average value of 64.5 msec). Stricter requirements increases the power consumption by an average of 32.8 % of the system with relaxed delay requirements (values are normalized by 7.9 kilowatt, i.e., consumption in fully distributed processing), whereas it improves the bandwidth consumption by an average of 38.6% (Normalized by 187.027 Gbps in fully centralized processing). Also, it is observed that the percentage of processing at CC in relaxed requirement ($FP@CC_R$ = 1 - $FP@EC_R$) is more than that of stricter case by an average of 32.56 %. This is realistic since pushing processing to the edge improve latency, given that processing capability at EC and CC is similar, $\eta_{EC} = 1.0$. Note that different efficiency values might result in different delay impacts on the power and bandwidth consumptions. Also, note that FP@EC does not reach zero, meaning that it is not possible to reach full centralization of functionals processing.

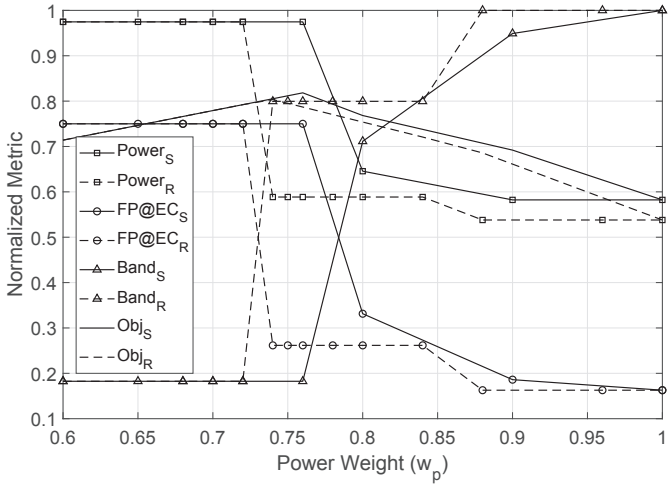

Fig. 4. Normalized Metrics, i.e., objective, power, and midhaul band versus weighting, w_p.

VI. Conclusion

In this work, we incorporated the end-to-end delay model, per user's request in the energy and bandwidth optimization framework. This model considers the impact of baseband function split options on the request delay, in a hybrid CRAN architecture. Our delay model takes 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, while highlighting the impact of different function split options. 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 system's power and bandwidth consumptions. Via numerical evaluation, it is found that the behavior of the end-to-end delay is highly impacted by the amount of processing at the edge and center clouds and the efficiency ratio between their equipment's processing power. We also found that stringent delay requirements (uniformly distributed around 45 msec for a content of 1Mbit size) increases power consumption by an average of 33 percentage, compared to relaxed delay uniformly distributed around 65 msec. As future work, we will consider a time-efficient heuristic algorithm to solve this problem, since the optimal solution of constraint programming model is complex and takes time as the size of the problem increases.

References

[1] C. M. R. Institute, "C-RAN: The Road Towards Green RAN," *Technical Report, Oct.*, 2011.

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

[3] C.-Y. Chang, N. Nikaein, R. Knopp, T. Spyropoulos, and S. S. Kumar, " FlexCRAN: A Flexible Functional Split Framework over Ethernet Fronthaul in Cloud-RAN ," in *To appear in proceedings of IEEE International Conference on Communication (ICC)*, 2017.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] A. Alabbasi and C. Cavdar, "Delay-aware green hybrid cran," in *2017 15th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, May 2017, pp. 1–7.

[10] X. Wang, A. Alabbasi, and C. Cavdar, "Interplay of energy and bandwidth consumption in cran with optimal function split," in *2017 IEEE International Conference on Communications (ICC)*, May 2017, pp. 1–6.

[11] 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.

[12] "40-Gigabit-capable passive optical networks (NG-PON2)," ITU-T G.989 series of Recommendations:, ITU-T, March 2013.

[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] "IBM ILOG CPLEX optimization studio: OPL language users manual." Version 12 Release 6, IBM, 2015.

[16] 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.

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