This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2018.2802419, IEEE Transactions on Green Communications and Networking

Optimal Processing Allocation to Minimize Energy and Bandwidth Consumption in Hybrid CRAN

*Abdulrahman Alabbasi, †Xinbo Wang, and *Cicek Cavdar
*Communication Systems Department, KTH Royal Institute of Technology, Sweden
†Computer Science Department, University of California Davis, USA
Email: *{alabbasi,cavdar}@kth.se, †xbwang@ucdavis.edu

***Abstract*—Cloud radio access network (CRAN) architecture is proposed to save energy, facilitate coordination between radio units (RUs), and achieve scalable solutions to improve radio network's performance. However, stringent delay and bandwidth constraints are incurred by fronthaul in CRAN (the network segment connecting RUs and digital units (DUs)). Therefore, we propose a hybrid cloud radio access network (H-CRAN) architecture, where a DU's functionalities can be virtualized and split at several conceivable points. Each split option results in two-level deployment of the processing functions (central site (CS) level and remote site (RS) level) connected by a transport network, called midhaul. We study the interplay of energy efficiency and midhaul bandwidth consumption under optimal processing allocation. We jointly minimize the power and midhaul bandwidth consumption in H-CRAN, while satisfying network's constraints, i.e., processing and midhaul bandwidth capacity. We enable power saving functionalities by shutting down different network's components. The proposed model is formulated as constraint programming problem. The proposed solution shows that 42 percentile of midhaul bandwidth savings can be achieved compared to the fully centralized CRAN; and 35 percentile of power consumption saving can be achieved compared to the case where all the network functions are distributed at the edge.**

I. Introduction

5G networks are envisioned to support 1000-fold more traffic and 10-fold lower latency. However, the cost and energy consumption should be sustainable [1]. In recent years, cloud radio access network (CRAN) has been proposed as a solution to reduce the power consumption and cost [2], [3]. In CRAN, DUs are decoupled from RUs and centralized at a CS, called DU hotel. A DU hotel is evolved to a DU cloud if general-purpose servers are employed in the DUs. Hence, their functions can be virtualized, which leads to virtualized CRAN. In this paper, we refer to the DU cloud as a central cloud. The centralization of DUs enables the sharing of not only computational resources but also the infrastructure in the central cloud. The authors of [4] studied how to save power consumption in a virtualized CRAN architecture. However, if all the baseband processing functions are centralized in the central cloud, midhaul bandwidth and transmission delay requirement become very challenging to satisfy by the link between RU and DU, called "fronthaul" (where I/Q samples generated by RUs must be transmitted to the central cloud). The 3ms delay constraint limits the distance between a central cloud and RU to 20~40 km. Also, a tremendous amount of bandwidth is required in fronthaul. For example, a single RU, with a 20 MHz carrier and 2x2 MIMO scheme, will generate 2.5 Gbps I/Q samples in downstream. In a densified radio access network (RAN) envisioned for 5G, such a rigid delay and bandwidth requirement leaves dedicated fiber or (active) optical transport network (OTN) as the only viable solution, which reduces the cost saving of CRAN.

This study is supported by EU Celtic Plus Project SooGREEN: Service-oriented optimization of Green mobile networks

The concept of splitting the communication functions processing has been recently proposed to overcome many of CRAN's drawbacks [5]. This technique divides the processing of communication functions into dual-sites processing instead of single-site processing. That is, DU's processing functions are split into two sites, the part close to RU is RS (called later as edge cloud) for partial communication processing, whereas the other part is placed in CS (called then as the central cloud) for the remaining processing. The processing functions, which are used for communication purposes, are classified into cell related processing and user associated processing as follows. The cell's processing functions are mainly connected to physical layer. (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 related operations and synchronization operation, (4) Resource mapping, (5) Channel estimation and equalization. The user's functions are related to physical, media access control (MAC), radio link control (RLC), and packet data convergence protocol (PDCP). (1) Channel estimation and equalization, (2) Modulation/demodulation MIMO (de)mapping and (pre)coding, (3) Forward error correction, turbo decoding, (4) RLC and PDCP function.

In this work, we call the transport network between RS and CS as "midhaul" [6], whereas, the link between RS and RU is referred to as fronthaul. We also assume that a single RS controls many RUs, hence, it is possible to implement coordinated multiple point techniques, e.g., joint transmission, among these RUs. The motivation of dual-site processing is twofold, a) by conducting partial communication processing at RS, bandwidth requirement can be significantly relaxed for midhaul; b) by equipping RSs with general-purpose processors, to support computational and content-caching capabilities, traffic load can be terminated at RS to relieve traffic load in core networks. In this study, we call CRAN with dual-site processing, at RS and CS, and function virtualization as hybrid cloud radio access network (H-CRAN). We assume that

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2018.2802419, IEEE Transactions on Green Communications and Networking

an RS is connected to user equipments (UEs) via RU that does not have processing capabilities. Hence, H-CRAN comprises the benefits of central and edge processing when needed.

In the context of realizing a practical cloud RAN, several works have been conducted, which can be classified as follows. Energy and cost studies of fronthaul, processing allocation, and processing split [7]–[9]. Others have investigated the design of a fronthaul protocol to deliver synchronized control and users' data information (which enables processing at two sites) from a center site (called baseband processing unit (BBU)) to remote radio head [10], [11]. Analytical studies, including market sharing (bids and asks), on techniques to enhance overall CRAN system's capacity (radio and processing) can be found in [12]–[15].

In [7], the authors have studied the impact of individual split on bandwidth and delay requirements of the fronthaul link, while considering several connectivity technologies, such as free space optic, DSL, millimeter wave, microwave, fiber (with different access), for the fronthaul link. Authors of [8] have considered a graph-based framework to reduce the cost of baseband processing and fronthaul via optimally placing the DU in the network. Dual-site processing is also considered in this work, where the remote site is connected to the remote unit or cell site. The authors of [8] interpreted the baseband transceiver as direct graphs so that the splitting and placement can be formulated as graph clustering problem. This problem is solved afterward via a genetic algorithm approach. Authors of [9] evaluated the impact of certain function splits, i.e., Physical, MAC, and PDCP-RLC splits, on the energy and cost savings of the CRAN network. Teletraffic theory has been used in the aforementioned quantitative study. This study considered a two-sites processing architecture with the remote sites being the base stations.

The authors of [10] have proposed several design requirements for the fronthaul to enable transmission of intermediate processing information (signaling and users' data) between the central site and base stations. Authors of [10] designed a fronthaul protocol to enable handling various traffic load, flexible topology, support different latency requirements. The authors of [11] studied reusing existing packet-based network (e.g., Ethernet) to decrease deployment costs of fronthaul of CRAN and cost of Baseband Unit (BBU) resources. Accurate phase and frequency synchronization impose a challenge in packet-based fronthaul. The authors of [11] 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 [12] have formulated a coexisting framework between several mobile network operators (MNOs) to share the CRAN's resources (physical resource blocks (PRBs) and number of BBUs). Their objective was to evaluate how the cooperation between MNOs can improve the social profit and quality of service (QoS). The problem is modeled as a coalition game and three MNOs were considered. Authors of [13] have considered the concept of physical resource transfer, in virtualized two-tier RAN, to improve the overall capacity. The concept of physical resource transfer is defined as the possibility of reconfiguring the orthogonal frequency-division multiple access (OFDMA)-based medium access of two base stations (BSs) to allow a BS to use a set of sub-carriers initially allocated to another BS. Moreover, authors of [14] focused on developing an analytical framework to evaluate the performance of Long-Term Evolution (LTE) enhanced by Fiber wireless. They addressed end-to-end delay (including backhaul), aggregated throughput, and wireless local area network (WiFi) offloading efficiency. Motivated by the benefits of turning off BSs, authors of [15] have introduced an offloading mechanism, where the operators lease capacity of a small cell network owned by a third party, to be able to switch off their BSs and maximize their energy efficiency when the traffic demand is low. The authors of [15] have devised a bidding strategy, truthful and rational auction scheme, and a multi-objective framework (which maximizes the profits of all parties and social energy consumption).

Unlike our work, none of the above works has evaluated the trade-off between the mobile network's energy and bandwidth consumptions via considering the impact of the optimal functional splits.

Deciding the optimal functional split is still an open problem, which profoundly depends on the objective to be optimized. Intuitively, if more functions are centralized at CS, as in CRAN case, higher power consumption saving can be achieved, whereas, midhaul bandwidth consumption will increase. On the contrary, placing more processing functions at the edge cloud may lead to higher power consumption but lower midhaul bandwidth consumption. Hence, a trade-off between placing the functions at CS or at RSs should be investigated.

In this study, our contributions are summarized as follows. First, we present the architecture of H-CRAN with midhaul and fronthaul links, and dual-site processing. In H-CRAN, DUs are deployed at both RSs and CS for baseband processing, whereas, each RS is connected to a group of RUs, see Sec. II. Second, we model the baseband processing chain as a sequence of functions that can be split between any two functions, and these functions can be deployed either in DUs at RSs or in DUs at CS. Third, we develop a mathematical model to solve this multi-objective problem using constraint programming to decide the optimal functional split per user in each RU, while aiming to jointly minimize the system's power consumption and the bandwidth consumption in midhaul. Fourth, we enable DU to shut down when users' load is low enough, i.e., current users can be processed by fewer number of DUs, compared with maximum load. Also, shutting down the cooling equipment at a remote site is possible given that no digital unit is active at that location. Another power saving feature is enabled by shutting down the fronthaul (millimeter wave link's components) and the remote unit (analog components and power amplifier) at very low load where the associated sites do not have any active user. Fifth, we evaluated the system performance by comparing its power consumption to benchmark systems, i.e., distributed-CRAN[1] and centralized-CRAN[2]. Finally, we evaluate the system's power and bandwidth consumption against different

[1]All baseband processing are distributed at the edge cloud.

[2]All baseband processing are centralized.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2018.2802419, IEEE Transactions on Green Communications and Networking

load ratio, number of RSs, and weighting parameter. We conclude that we can achieve considerable saving in both power and bandwidth consumptions, using the optimal placement of baseband processing functions in H-CRAN architecture. By joint optimization and partial centralization of the functions, 42% midhaul bandwidth savings can be achieved compared to the fully centralized C-RAN solution; and 35% power consumption saving can be achieved compared to the case where all the network functions are distributed at the edge. These results demonstrate a compromise in between two extremes. Preliminary results of this study are published in [16].

The organization of the remaining sections is as follows. Section II addresses the architecture of H-CRAN and presents the functional split model. Section III presents the problem formulation. Finally, Sec. IV presents the simulation results.

II. Network Architecture

A. Proposed H-CRAN Architecture

We present a hybrid architecture that employs dual-site processing in virtualized CRAN. In this architecture, DUs are deployed at both CS in the central cloud and RS in edge cloud. So that baseband processing can be flexibly provisioned by a chain of virtualized functions for a RU or even for an UE during the transmission of user's traffic, end-to-end from the UE connected to an RU to the centralized cloud. 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 a RU is referred to as a "cell"), RS layer at edge cloud, and CS layer at the central cloud. Cell layer consists of cells that are being densified, each serving several UEs. A group of cells is connected to an RS as an aggregation point. The fronthaul between a cell and an RS is implemented using mmWave links [17] [3]. The RSs can be connected to CS via midhaul using various technologies, from expensive dark fiber or TON solutions to cost-efficient PON families or other Ethernet-based technologies. In this work, we study the system's power consumption under the constraint of bandwidth capacity per midhaul link (for an RS), which ranges from 1 Gbps to 20 Gbps. The midhaul technology considered in this study is time-wavelength division multiplexing PON (TWDM-PON) [18], and each midhual link is a wavelength channel, which needs an optical network unit (ONU) at RS and a Line-Card (LC) at CS as transceivers.

Edge cloud (RS) layer and central cloud (CS) layer are deployed with DUs, which are containers for virtualized functions because their computational resources can be virtualized and shared by any connected RU (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 the central cloud. However, RS is usually less energy-efficient than CS, because the number of DUs, associated with RUs, at the CS is larger than that in each RS. Hence, sharing infrastructure equipment, such as cooling, results in higher energy saving at CS. The trade-off becomes whether to centralize functions at RS in edge cloud layer (to save midhaul bandwidth), or to centralize more functions at CS in central cloud layer (to save power).

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

- We enable DUs to shut down when users' processing load at the remote site is low enough, i.e., users associated with the remote site can be processed by fewer number of DUs[4], compared with maximum load. This feature is also enabled for the CS's DUs.
- We enable shutting down the cooling equipment at RSs or CS given that no digital unit is active at that site.
- We enable shutting down midhaul, fronthaul, and radio access components at very low load where some remote sites do not have any active user.
- We enable shutting down a wavelength, given that the associated cells do not have any active user, i.e., this remote site does not consume any midhaul bandwidth. This feature can be triggered when the number of cells per remote site is low, or the system load is very low. Hence, it is possible to shut down the whole wavelength of that specific remote site.

B. Reference Architectures

We define two benchmark cases with no functional split as reference architectures for the performance analysis. (1) Edge-CRAN where all the baseband functions are kept at the edge cloud (distributed among several RSs) and the connection to the central cloud is provided by a backhaul. In this case, DUs are stacked at a nearby cabinet within the RS to serve RU of the BS, and cannot be shared by other BSs, which leads to low energy efficiency. But since baseband processing is conducted at RS, 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 [19]. However, since all processing is conducted at CS, a massive amount of bandwidth is needed in the fronthaul.

C. Functional Split Model

To study the distributed vs. centralized processing of network functions, we model the functional split of baseband processing chain for a cell, as shown in Fig. 2. First, baseband processing for a cell is modeled as a chain of functions, which includes m Cell-Processing (CP) functions and n User-Processing (UP) functions[5]. 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 include serial-to-parallel conversion (or common public radio interface (CPRI) encoding), removing the cyclic

[3]Our architecture is not hard-wired to any specific fronthaul technology within RS, as it is not the focus of our study.

[4]This occurs either because the number of users is extremely low or the users' processing is conducted at the central site.

[5]In this study, $m = 3$ and $n = 3$, as considered in [5], [20]–[22].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2018.2802419, IEEE Transactions on Green Communications and Networking

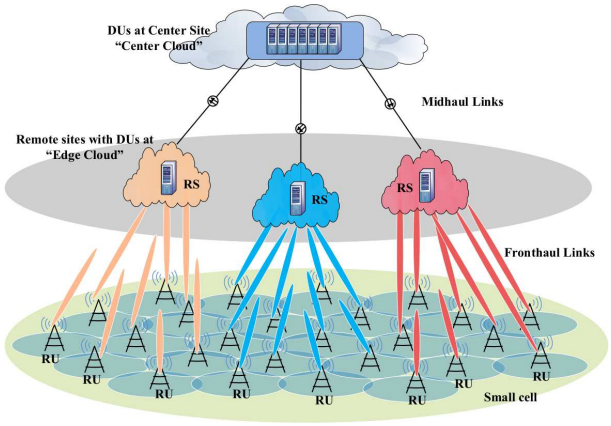

Fig. 1. Hybrid virtualized RAN architecture.

prefix, fast Fourier transform, and finally resource demapping, etc. The per-cell processing will be terminated at CP_m, and now signals from a cell will be de-multiplexed as multiple signal streams, each belonging to an UE. Then, UPs are the sequence of functions that will continue to process the signal streams on a per-UE basis[6], including equalization, inverse discrete Fourier transform, quadrature amplitude modulation, antenna demapping, multi-antenna processing, forward error correction, turbo decoding, and other RLC and PDCP functions.

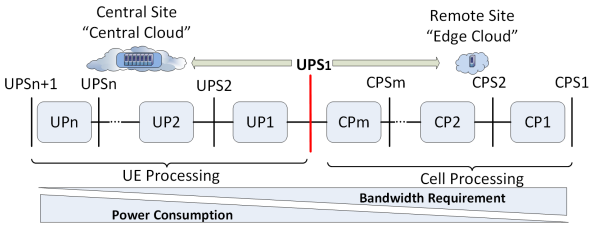

Fig. 2. Function split model.

As shown in Fig. 2, a 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 distributed radio access network (DRAN), characterized by a fully distributed deployment.

In Fig. 3, we further illustrate our assumptions on the implementation of a functional split in H-CRAN. First, the processing chain can be cut at most once and thus split into two parts, and the lower parts will be deployed at RS, and the upper parts will be placed at CS. Second, each functional split will incur a bandwidth requirement, which can be calculated by formulas provided in [5]. Third, once a functional split is decided, we can determine the number of CPs and UPs that need to be deployed at RS and CS, respectively. For example, in the left part of Fig. 3, for $cell_\alpha$, split happens in CP sequence, so massive amount of bandwidth is required to transmit the partially-processed signals. Hence, all CPs below the split point must be placed at the same DU in RS, and all CPs above the split point and all UPs must be placed at

[6]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.

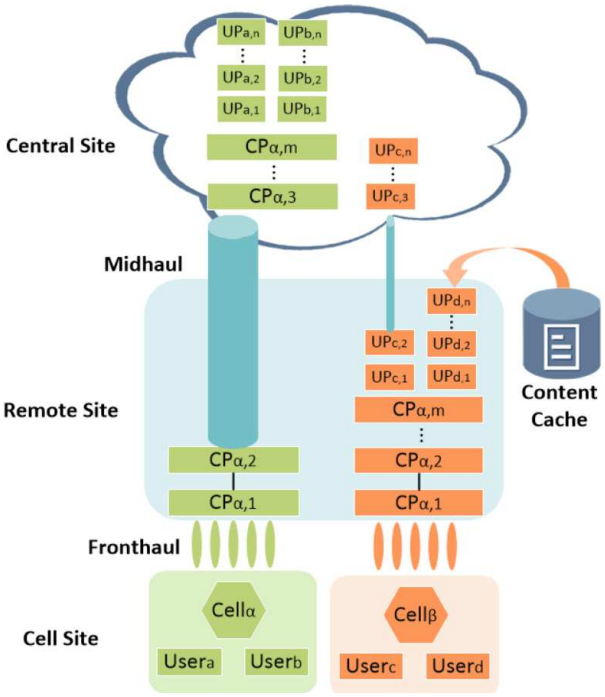

Fig. 3. H-CRAN including functional split architecture.

the same DU in CS. However, for $cell_\beta$, split happens in UP sequence, as shown in the right part of Fig. 3. Hence, larger amounts of cell's and users' signals have been processed so that less bandwidth will be required in midhaul. Also, since cell's signals have been de-multiplexed, different functional splits can be conducted and customized for different UEs. For example, for a UE that needs latency-critical service, UP should be split at a higher point, resulting in more processing at the edge cloud. If content caches are deployed at RS, latency-critical service can be provided by deploying all UPs at the edge cloud, RS. UE based service differentiated cloud resource provisioning is left as a future work in this paper.

III. Functional Split Optimization Problem

In this section, we aim at jointly minimizing the system's power consumption and bandwidth consumption of midhaul for H-CRAN. Therefore, we model the interplay of power and bandwidth consumptions as a functional split optimization problem, which is formulated as a constraint programming problem [23].

A. Given

- Topology: one CS connected to multiple RSs. Each RS is connected to an exclusive set of cells, and each cell exclusively covers a set of UEs
- $\mathbb{I}_x$: set of UEs. When $x = 0$, it refers to all UEs in H-CRAN, otherwise, it refers to set of UEs in cell $x = c$.
- $\mathbb{C}_x$: a set of cells. When $x = 0$, it refers to all cells in H-CRAN, $x = r$ refers to set of cells belonging to RS r.
- $\mathbb{D}_x$: a set of DUs. When $x = 0$, it refers to all DUs in H-CRAN, $x = -1$ refers to set of DUs in the CS , $x = r$ refers to set of DUs in RS r.
- $\mathbb{R}$: a set of all RSs.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2018.2802419, IEEE Transactions on Green Communications and Networking

- $\mathbb{W}$: a set of wavelengths.
- F_x: a set of functional split options, where x represents UP split or CP split.
- $H_x^y(.)$: pre-calculated mapping from a split option $x = \{UP, CP\}$ to the number of (UP and CP) functions at site $y = \{CS, RS\}$. For example, if the UP sequence of a UE is split at the middle point (indexed by m), $H_{UP}^{RS}(m) = H_{UP}^{CS}(m)$ equals to half of UPs.
- $J_i(.)$: pre-calculated mapping from UP split of UE i to the required midhaul bandwidth, which is proportional to the number of RBs allocated to UE i [5].
- $G_c(.)$: pre-calculated mapping from CP split of cell c to the required midhaul bandwidth, which is proportional to the number of antennas and carrier bandwidth [5].
- K: bandwidth capacity of a wavelength. Note that this is different from bandwidth induced and consumed by user's and cell's processing split, described with $J_i(.)$ and $G_c(.)$.
- L_x^y: the capacity of a DU located at the "y" site, $y = \{CS, RS\}$, in terms of the number of x functions that can be accommodated by this DU (x represents CP or UP, and y represents CS or RS). For example, L_{CP}^{CS} represents the maximum number of CPs that can be accommodated by a DU at CS.
- We also define an indicator function, $\mathcal{I}(.)$, to test whether a constraint is satisfied. If the constraint in the argument of the function, is satisfied, the indicator function has value 1, otherwise, 0. The mathematical definition is expressed as follows,

$$\mathcal{I}(a = b) = \begin{cases} 1; & \text{if} \quad a = b, \\ 0; & \text{if} \quad a \neq b. \end{cases} \tag{1}$$

B. Integer Variables

- $p_i \in [0, |F_{UP}|]$: denotes the UP functions split of UE i. Higher p_i means that we distribute a larger number of UP functions at RS (because we consider all functions processing are originally centralized as in CRAN), hence, if $p_i = |F_{UP}|$ then all UP functions are distributed, otherwise, if $p_i = 0$ then all UP functions are centralized.
- $q_c \in [0, |F_{CP}|]$: denotes the CP functions split of cell c. Higher q_c means that we distribute a larger number of CP functions at RS, hence, if $q_c = |F_{CP}|$ then all CP functions are distributed, otherwise, if $q_c = 0$ then all CP functions are centralized.
- $m_i \in D_r$: DU hosting UPs of UE i at RS r. Note that since the association between i and r is fixed, UE i can only choose a DU from a given set.
- $n_i \in D_{-1}$: DU hosting UPs of UE i at CS.
- $x_c \in D_r$: DU hosting CPs of cell c at RS r.
- $y_c \in D_{-1}$: DU hosting CPs of cell c at CS.
- w_r: wavelength used by RS r.
- l_r: number of active DUs at RS r.
- l: number of active DUs at CS.
- g: number of active wavelengths in the midhaul.

C. Objective

Our objective is to minimize a linearly weighted sum of the system's power consumption plus the total bandwidth consumption in midhaul. The power and bandwidth consumptions are normalized. The multi-objective function is expressed as,

$$\min w_P \cdot \frac{\mathcal{P}_T}{p_N} + w_B \cdot \frac{\mathcal{B}_{MH}}{b_N}, \tag{2}$$

where w_P is the weighting factor of the power consumption, and w_B is the weighting factor of the midhaul bandwidth consumption. The parameters p_N and b_N are the normalization factor of each the power and bandwidth consumptions, respectively[7]. We choose w_P = 1 - w_B to highlight the complementary impact of optimizing both bandwidth and power consumptions on the allocation of the processing functions. The notations $\mathcal{P}_T$ and $\mathcal{B}_{MH}$ denote the total power consumption[8] and midhaul bandwidth consumption, respectively. The total power consumption is expressed as,

$$\begin{aligned} \mathcal{P}_T = & \left(P_{CS} + lP_{CS}^{DU}\right)\mathcal{I}\,(l > 0) + gP_{lc} + \\ & \sum_{r\in\mathbb{R}}\left[\left(\sum_{c\in\mathbb{C}_r}(P_{Tx} + P_{FH})\mathcal{I}(|\mathbb{I}_c| > 0)\right)\right. \\ & \left. + \left(\mathcal{I}\left(\sum_{c\in\mathbb{C}_r}|\mathbb{I}_c| > 0\right)P_{onu} + P_{RS}\mathcal{I}\,(l_r > 0) + l_r P_{RS}^{DU}\right)\right], \end{aligned} \tag{3}$$

where the power consumption of DU at CS and RS are expressed as P_{CS}^{DU} and P_{RS}^{DU}, respectively. The power consumption of LC, ONU, fronthaul and radio links transmissions, housing at both CS and RS are expressed respectively as P_{lc}, P_{onu}, P_{FH}, P_{Tx}, P_{CS}, and P_{RS}. The parameters l and l_r are the number of active DUs at CS and r^{th} RS (where the integer $r \in \{0, ..., |\mathbb{R}|\}$), 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 RSs, i.e., $r \in \{0, ..., |\mathbb{R}|\}$ and the associated cells, $c \in \{0, ..., |\mathbb{C}|\}$, as follows,

$$\mathcal{B}_{MH} = \sum_{w\in\mathbb{W}}\sum_{r\in\mathbb{R}}\mathcal{I}\,(w_r = w)\sum_{c\in\mathbb{C}_r}\left(G_c(q_c)\mathcal{I}(|\mathbb{I}_c| > 0) + \sum_{i\in\mathbb{I}_c}J_i(p_i)\right), \tag{4}$$

where $G_c(q_c)$ is a function that relates q_c to the required midhaul bandwidth [5]. The function $J_i(p_i)$ relates the user processing split, p_u (of the i's user, where $i \in \{0, ..., |\mathbb{I}_c|\}$), to the required midhaul bandwidth, can be found in [5]. The term $\mathcal{I}\,(w_r = w)$ ensures that the current wavelength belongs to the remote site r.

The calculation of overall power, $\mathcal{P}_T$ in (3), clearly describes the functionalities of power saving that has been mentioned in the previous section. The terms $\left(P_{CS} + lP_{CS}^{DU}\right)\mathcal{I}\,(l > 0)$ and $P_{RS}\mathcal{I}\,(l_r > 0) + l_r P_{RS}^{DU}$ represent shutting down the site if there are no active DUs. Whereas, terms lP_{CS}^{DU} and $l_r P_{RS}^{DU}$ represent shutting down the inactive DUs. The terms $\mathcal{I}\left(\sum_{c\in\mathbb{C}_r}|\mathbb{I}_c| > 0\right)P_{onu}$ and gP_{lc} represent shutting down the LC and ONU if there are no active users in the associated RS. Finally, the term $\sum_{c\in\mathbb{C}_r}(P_{Tx} + P_{FH})\mathcal{I}(|\mathbb{I}_c| > 0)$ represents shutting down the fronthaul and radio access transmission components if there are no active users in the associated cell.

[7]The normalization factors, p_N and b_N are the maximum consumed power and maximum consumed midhaul bandwidth, respectively.

[8]The total consumption of the whole network, from CS to the cells

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2018.2802419, IEEE Transactions on Green Communications and Networking

Expression (4) describes shutting down the wavelength of a certain remote site, i.e., no active users exist in this RS. To explain when this situation occurs, we need to define the number of active users per number of remote sites ratio, i.e., $\frac{|\mathbb{I}_0|}{|\mathbb{R}|}$. Note that this ratio becomes very low when we have a high number of remote sites, i.e., $|\mathbb{R}|$ is large, and we operate in very low load, low $|\mathbb{I}_0|$. In these conditions, it is with high probability to have a remote site with no active users under its corresponding cells. Hence, the associated wavelengths can be shut down.

D. Constraints

In this sub-section, we explain the constraints of the problem.

$$\mathcal{I}(p_i < |F_{UP}|) + \mathcal{I}(q_c < |F_{CP}|) = 1, \qquad \forall i \in \mathbb{I}_c,\ \forall c \in \mathbb{C}_0. \tag{5}$$

Constraint (5) ensures that function split can occur only once, either at CP or UP.

$$(p_i < |F_{UP}|) \implies (m_i = x_c), \qquad \forall i \in \mathbb{I}_c,\ \forall c \in \mathbb{C}_0. \tag{6}$$

Constraint (6) ensures that if UP of UE i is split, then lower part UPs must be placed in the same DU with their CP at RS, as shown in Fig. 2, because otherwise complex inter-DU communication will be incurred.

$$(q_c < |F_{CP}|) \implies (n_i = y_c), \qquad \forall i \in \mathbb{I}_c, \forall c \in \mathbb{C}_0. \tag{7}$$

Constraint (7) ensures that if CP of cell c is split, the upper part of CPs must be placed in the same DU with all UPs (of all UEs in cell c) at CS, as shown in Fig. 2.

$$\sum_{c \in \mathbb{C}_r} H^{RS}_{CP}(q_c) \cdot \mathcal{I}(x_c = d) \le L^{RS}_{CP}, \qquad \forall r \in \mathbb{R},\ \forall d \in \mathbb{D}_r. \tag{8}$$

Constraint (8) ensures that the total number of CPs that are accommodated by a DU d at RS r cannot exceed this RS-DU's CP capacity. Note that L^{RS}_{CP} is less than L^{CS}_{CP}.

$$\sum_{c \in \mathbb{C}_0} H^{CS}_{CP}(q_c) \cdot \mathcal{I}(y_c = d) \le L^{CS}_{CP}, \qquad \forall d \in \mathbb{D}_{-1}. \tag{9}$$

Constraint (9) ensures that the number of CPs that are accommodated by a DU d in CS cannot exceed this CS-DU′s CP capacity.

$$\sum_{c \in \mathbb{C}_r} \sum_{i \in \mathbb{I}_c} H^{RS}_{UP}(p_i) \cdot \mathcal{I}(m_i = d) \le L^{RS}_{UP}, \qquad \forall r \in \mathbb{R}, \forall d \in \mathbb{D}_r. \tag{10}$$

Constraint (10) ensures that the number of UPs that are accommodated by a DU d at RS r cannot exceed this RS-DU's UP capacity.

$$\sum_{i \in \mathbb{I}_0} H^{CS}_{UP}(p_i) \cdot \mathcal{I}(n_i = d) \le L^{CS}_{UP}, \qquad \forall d \in \mathbb{D}_{-1}. \tag{11}$$

Constraint (11) ensures that the number of UPs that are accommodated by a DU d at CS cannot exceed this CS-DU's UP capacity.

$$\sum_{r \in \mathbb{R}} \mathcal{I}(w_r = w) \cdot \sum_{c \in \mathbb{C}_r} \left(G_c(q_c) + \sum_{i \in \mathbb{I}_c} J_i(p_i) \right) \le K, \qquad \forall w \in \mathbb{W}. \tag{12}$$

Constraint (12) ensures that the total occupied midhaul bandwidth in a wavelength (given by left-hand side of (12)) cannot exceed the wavelength's capacity, i.e., K. The occupied bandwidth in wavelength w is the sum of the bandwidth consumptions of all RSs that are using w. The bandwidth consumption of RS r is the sum of bandwidth consumptions of all cells belong to it ($c \in \mathbb{C}_r$). The bandwidth consumption of cell c is either the bandwidth requirement incurred by CP split ($G_c(q_c)$), or the bandwidth requirement incurred by UP splits. If it is the later, the bandwidth requirement of cell c is the sum of bandwidth requirements of all UEs ($J_i(p_i)$) belonging to c ($i \in \mathbb{I}_c$).

$$l_r = \sum_{c \in \mathbb{C}_r} \delta(x_c), \quad \forall r \in \mathbb{R}, \tag{13}$$

where $\delta(x_c) = 0$, if x_c is not active (no CP of cell c is placed at RS r), $\delta(x_c) = 1$, if x_c is active (at least one CP of c is placed at RS r). Constraint (13) counts the number of active DUs at RS r. The algorithmic model of this constraint is expressed as,

$$l_r = \text{countDiff}\left(\{x_c\}_{c \in \mathbb{C}_r}\right) - \prod_{c \in \mathbb{C}_r} (q_c = 0), \quad \forall r \in \mathbb{R}. \tag{14}$$

The countDiff operator counts the number of distinct values taken by variables in array $\{x_c\}_{c \in \mathbb{C}_r}$. When there exist active DUs in RS r, the number of active DUs is equal to the number of distinct values taken by $\{x_c\}_{c \in \mathbb{C}_r}$. But in the special case when there are no active DUs in RS r, all cells of the RS r choose the rightmost split in Fig. 2, i.e. $q_c = 0, \forall c \in \mathbb{C}_r$, so $\prod_{c \in \mathbb{C}_r} (q_c = 0)$ equals to 1. To ensure that $l_r = 0$ in this case, all x_c are forced to choose the same value, i.e., countDiff $\left(\{x_c\}_{c \in \mathbb{C}_r}\right)$ equals to 1.

$$l = \text{countDiff}\left(\{n_i\}_{i \in \mathbb{I}_0}\right) - \prod_{i \in \mathbb{I}_0} (p_i = |F_{UP}| - 1). \tag{15}$$

Constraint (15) counts the number of active DUs at CS. The explanation of (14) can apply to this constraint also, except that the special case that there are no active DUs at CS happens when all UEs choose the leftmost split in Fig. 2, i.e. $p_i = |F_{UP}| - 1, \forall i \in \mathbb{I}_0$.

$$g = \text{countDiff}\left(\{w_r\}_{r \in \mathbb{R}}\right). \tag{16}$$

Constraint (16) counts the number of midhaul's active wavelengths.

IV. Simulation Results

In this section, we evaluate the system performance of the multi-objective function split optimization and analyze the interplay between power and midhaul bandwidth consumption as

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2018.2802419, IEEE Transactions on Green Communications and Networking

functions are centralized at the central cloud versus distributed at the edge clouds. As a reference, we consider two cases to compare the optimized network function placement, (1) Edge-CRAN, where all the functions are placed at the edge cloud, i.e., at the remote site; and (2) Central-CRAN, where all the functions are centralized at the central cloud. We solve the proposed framework using IBM ILOG CP solver[9]. This solver obtains the optimal value if the parameter optimality tolerance gap is set to 0. However, the nature of the problem includes bin packing and therefore it is known to be an NP-hard problem which means the time complexity will grow exponentially with the problem size. By increasing the value of optimality gap, we have analyzed the scalability of the solution method as the problem size grows. The system parameters are expressed in Table I.

TABLE I
SIMULATION PARAMETERS.

Parameter Name	Value
Topology	1 CS, 6 RSs, 7 RUs per RS.
Configuration of RU[10]	20 MHz, 2*2 MIMO, 64 QAM (DL)
Number of CPs & UPs per RU	3 & 3
Capacity of DU at RS	3 CPs and 15 UPs (one RU)
Capacity of DU at CS	27 CPs and 135 UPs (nine RUs)
Power of DU at RS	50 W
Power of infrastructure RS	250 W
Power of DU at CS	100 W
Power of infrastructure CS	500 W
LC Power+ ONU Power,	20 W + 5 W
Radio Access + Fronthaul link Power Consumption [11].	20, 40 W
Users per RU	2, 4 [12]
K (Capacity limit of midhaul wavelength)	20000 Mbps

In Fig. 4, we study how function placement changes as more midhaul bandwidth is available, if we emphasize on minimizing the power consumption ($W_P = 0.99$ in Eqn. (2)). We plot the percentages of functions that are placed at RS and CS, respectively, versus the bandwidth capacity per midhaul link (wavelength). As bandwidth capacity increases, more functions are centralized at CS, and fewer functions are placed at RSs. Because when power consumption is minimized, and midhaul bandwidth is abundant, it is always preferable to centralize functions at an energy-efficient central cloud. Also, there exists an upper bound for effective usage of the wavelength, because when bandwidth is larger than 13 Gbps in Fig. 4, all functions have already been centralized, and thus no more power can be saved. However, in practice, not all

[9]ILOG CP solver addresses a vast range of discrete scheduling and combinatorial optimization problems. It efficiently solves the packing problems and enables the incorporation of logical constraints. One of the key features of constraint programming is the ability to exploit the structure of a problem to design an adaptive search strategy to solve it. CP optimizing engine adaptively uses many novel concepts to find the optimal solution efficiently [24]. For instance, it uses adaptive search, which consists of several heuristic search techniques, including large neighborhood search and genetic algorithms. The nature of our problem includes bin packing problem and combinatorial optimization. It also has several logical constraints. Since it aligns with the capabilities of CP tool, we have chosen CP to solve our problem.

[10]The configuration parameters are used to calculate the bandwidth consumption of a RU mapped from a functional split, using formula provided in [5], [25]

[11]Contains the power consumption of the connection link from RS to RUs and power radiated from RUs.

[12]We consider a fully loaded scenario, unless otherwise mentioned, where each RU assign all physical resource blocks to existing UEs.

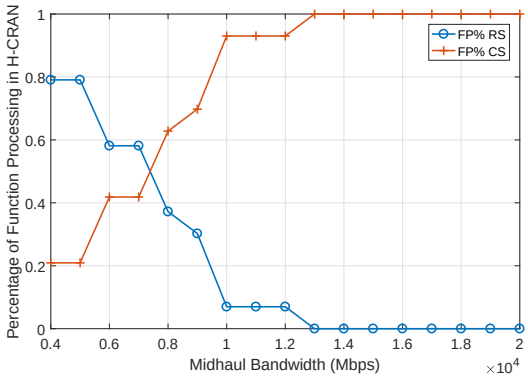

Fig. 4. Illustration of percentage of functions placed at RSs/CS, versus bandwidth capacity per midhaul link (under $w_P = 0.99$).

network operators can have sufficient midhaul bandwidth, so the planning of function placement is mutually decided by the available midhaul solutions (with different bandwidth capacity per link).

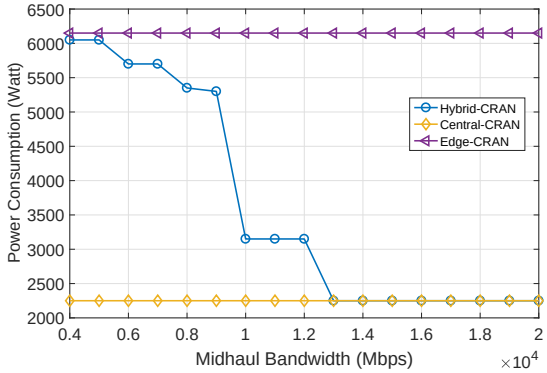

Fig. 5. Power consumption of hybrid-CRAN, central-CRAN, and edge CRAN versus midhaul bandwidth (under $w_P = 0.99$).

Fig. 5 compares the power consumption of the proposed H-CRAN with the power consumptions of central-CRAN and edge-CRAN, versus the bandwidth capacity per midhaul link, when power consumption is more valued ($w_P = 0.99$). Confirming the result in Fig. 4, as more bandwidth is available in midhaul, less power will be consumed by hybrid-CRAN. When bandwidth capacity is small, hybrid-H-CRAN achieves the same power consumption as edge-CRAN, because all functions are centralized at the edge cloud layer divided between RSs. When bandwidth capacity is large, H-CRAN achieves the same power consumption of central-CRAN, because all functions are centralized at CS.

Fig. 6 presents a comparison between different normalized system performances, i.e., the objective value, power consumption, midhaul bandwidth consumption, and percentages of functions placed at RSs and at CS, respectively. This evaluation is conducted for different w_P values [13]. When w_P increases, the power consumption decreases, and the bandwidth consumption increases, which indicates the trade-off between power and bandwidth consumptions. Moreover,

[13]The normalization factor for power is 4500Watts and for bandwidth is 105.7Gbps, which are the associated max value without an optimal solution in this setting.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2018.2802419, IEEE Transactions on Green Communications and Networking

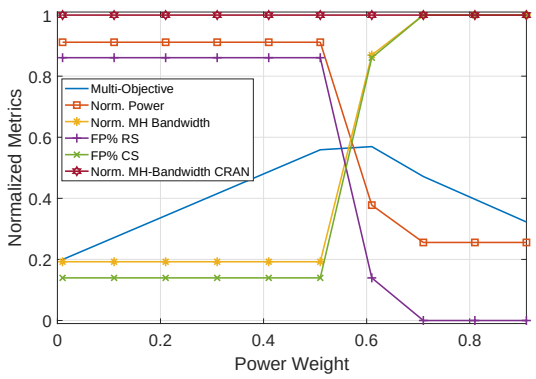

Fig. 6. Normalized metric, system performances, versus power weighting parameter.

the curves of power consumption and bandwidth consumption transpose when w_P ranges from 0.5 to 0.6, which indicates that their interplay is a drastic process. So, if we want to minimize power and bandwidth consumptions jointly, their weights in objective function must be carefully chosen, e.g., $w_P = 0.55$ where the two curves cross. At this optimal point, around 35% and 42% of savings can be achieved at power and bandwidth consumptions, respectively[14]. It is also observed that as there is more stress on power minimization, the bandwidth consumption of H-CRAN converges to that of central-CRAN, given that there is no constraint on midhaul bandwidth capacity.

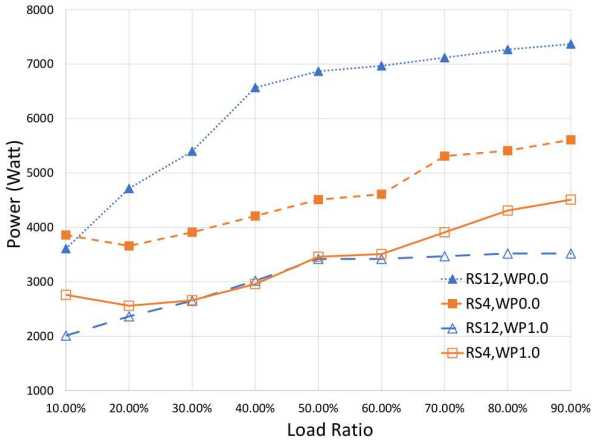

Fig. 7. Total Power consumption versus Load Ratio for different number of RSs.

Fig. 7 and Fig. 8 evaluate the total power and midhaul bandwith consumptions of the network versus load ratio for different number of RSs. Obviously, we note that power and bandwidth increase with the increase of load. One main conclusion is that high number of RSs achieves better performance, at high load, of the metric (power or bandwidth) that is associated with the maximum weighting, i.e., $w_P = 1$ for power and $w_P = 0$ for bandwidth. This occurs because the saving function works best at a high number of RSs given varying load ratios. It is also noted that the individual targeted metric to be optimized (power at $w_P = 1$ and bandwidth at

[14]This is in compared to distributing all processing at RSs (for power saving) and centralizing all processing at CS (for bandwidth saving).

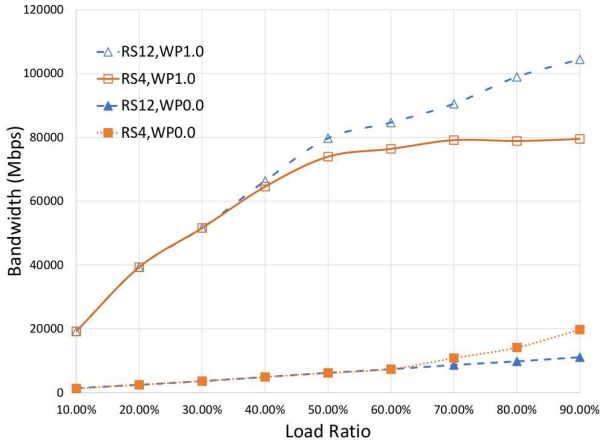

Fig. 8. Midhaul Bandwidth consumption versus Load Ratio for different number RSs.

$w_P = 0$) varies the least with the change of load. For instance, Fig. 8 shows the maximum variation of bandwidth is only 18.5 Mbps at $w_P = 0$, whereas the variation is about 85.2 Mbps at $w_P = 1$.

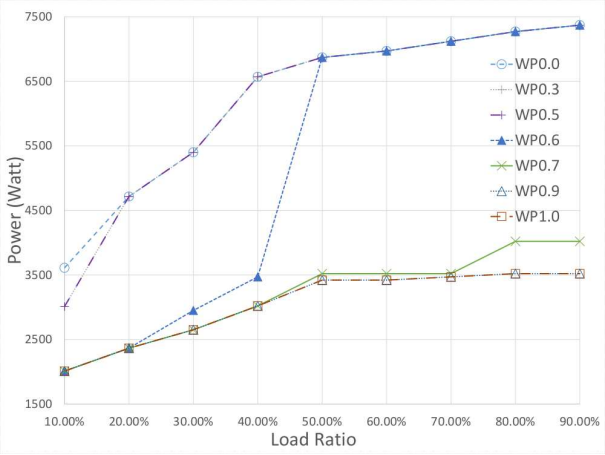

Fig. 9. Total Power versus Load Ratio for RS=12.

Fig. 9 evaluates the power performance versus load ratio. Intuitively with the increase of the load ratio, the power consumption increases. Load-dependent contributors to the power consumption figure are (1) radio frequency transmission power, (2) power of required DUs to conduct UPs and CPs, (3) power of midhaul and fronthaul links. Also, variable w_P, i.e., $\{0.0,0.3,0.5,0.6,0.7,0.9,1.0\}$, induces different behaviors of power consumption. For w_P=$\{0.0,0.3,0.5\}$, call it group 1, relatively high power consumption is achieved, where all of them have similar consumption for a large portion of load ratios. Whereas, the power consumptions of $\{0.7,0.9,1.0\}$, call it group 2, are similar in a portion of load ratio, and it is relatively low compared with group 1. At $w_P = \{0.6\}$, the power consumption changes with the load from the low consumption group (at low load ratio) to the high consumption group (at high load ratio) at around 50% load ratio. Hence, depending on our load, we can design w_P to achieve the desired performance. Also, the performance of w_P=0.5 starts at low load with better consumption compared to group 1 then have similar performance. The reason for this behavior is that

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2018.2802419, IEEE Transactions on Green Communications and Networking

the power minimization at w_P=0.5 is valued more than that at w_P=0.0. Another reason is at low load many DUs at different RSs are switched off [15] then switched on with the increase of load [16].

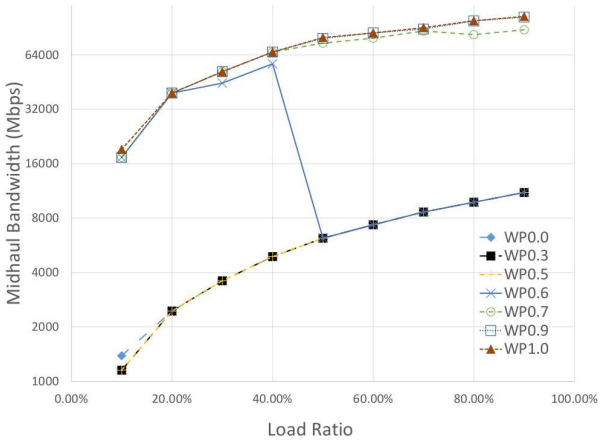

Fig. 10. Total Midhaul Bandwidth versus Load Ratio for RS=12.

Fig. 10 evaluates the midhaul bandwidth consumption versus load ratio[17]. We note that the weighting factor split the bandwidth consumption into two groups, i.e., {0.0,0.3,0.5} and {0.7,0.9,1.0}. The performance of group 1 (low w_P) is almost similar in large portion of load ratios, and it achieves relatively low bandwidth consumption. However, group 2 (high w_P) have high bandwidth consumption, and the consumption slightly varies at low and high load ratios. This is due to the allocation of a low number of users per cell (at low load), hence per RS (for a high number of RSs)[18]. Therefore, any slight increase in the load will impact the switching on wavelengths at the associated RS. We compared this result to RS = 4 case, where the performance is similar at high w_P group for low and high load ratios. Also, we note that at w_P=0.6 the performance changes from higher consumption to lower consumption with the increase of load ratio. This is because at medium w_P and high load ratio the consumption of midhaul link is high and comparable to that of DUs. Hence, allocating processing at RS saves both power and bandwidth, as shown in Fig. 9.

The main conclusion of both Fig. 9 and Fig. 10 is that based on the actual load of the network we can find the best w_P to optimize both power and bandwidth consumptions.

Fig. 11 evaluates the power consumption performance versus power weight, i.e., $w_p = 1 - w_b$, for different load ratio values, L = { 0.1, 0.5, 0.7, 0.9}. This evaluation is conducted for RS=12. Lower load ratios result in remarkably lower power consumption compared to medium to high loads, which perform slightly similar. This is because at very low load ratio, when RS=12, there is a considerable number of RSs that do not consume processing, transportation, and radio transmission;

[15]This occurs for many reasons, one of which is that the associated user/cell processing are allocated in CS instead of RS.

[16]To accommodate higher load processing.

[17]The bandwidth consumption results, shown in Fig. 10, are associated with the power consumption results in Fig. 9.

[18]This can be expressed as the ratio of users per remote site, $\frac{|\mathbb{U}_0|}{|\mathbb{R}|}$

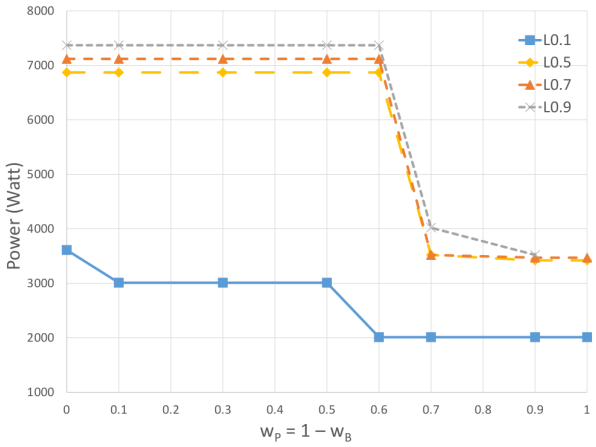

Fig. 11. Total Power versus $w_P = 1 - w_B$ for RS=12.

hence it can be switched off. At low load (L=0.1), the system consumes less power compared with high load (L=0.9) by about 1.51 kWatts to 3.76 kWatts, i.e., about 20.4 % to 51% (of high load consumption). However, since the system's load ratio is a parameter that cannot be controlled by the system designer, it is interesting to see the amount of possible saving given specific load ratios, i.e., {0.1,0.5,0.7,0.9}. Hence, we now evaluate how much saving is possible at specific load ratios by controlling w_P. It is observed that at low load the maximum possible saving of power consumption (by tuning w_P from w_P = 0 to w_P = 1) is only 1.6 kWatts, i.e., about 44.3%, while 3.8 kWatts, i.e., 52.2%, saving can be achieved at high load. That is, the possibility of saving (by tuning design parameter) in high load is two times more than that in low load (in Watts).

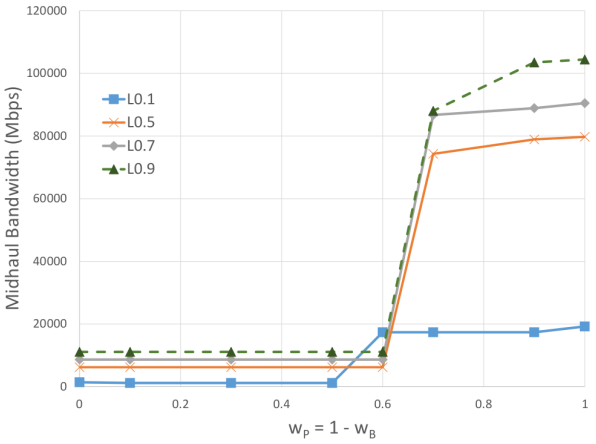

Fig. 12. Total Bandwidth versus $w_P = 1 - w_B$ for RS=12.

Fig. 12 evaluates the midhaul bandwidth consumption performance versus variable power weight, i.e., $w_p = 1 - w_b$, for different load values, L = { 0.1, 0.5, 0.7, 0.9}. This evaluation is for RS=12. It is observed that at low w_P value the variation in bandwidth consumption (for different load ratios) is not very large compared with that of large w_P. This is because at low w_P the optimization of bandwidth becomes dominating priority, hence minimum impact of load changing is observed. At low load (L=0.1), the system consumes less bandwidth

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2018.2802419, IEEE Transactions on Green Communications and Networking

compared with high load (L=0.9) by about 9.7 Gbps to 85.230 Gbps, i.e., about 9.3% to 81.6% (of high load consumption). In similar lines to Fig. 11, we check the impact of tuning system design parameter (w_P), at specific load, on bandwidth consumption. At low load (L=0.1) the maximum possible saving of bandwidth consumption (when tuning w_P) is about 17.8 Gbps (≈92% bandwidth savings at L=0.1), while 93.3 Gbps saving can be achieved at high load (≈89% bandwidth savings at L=0.9).

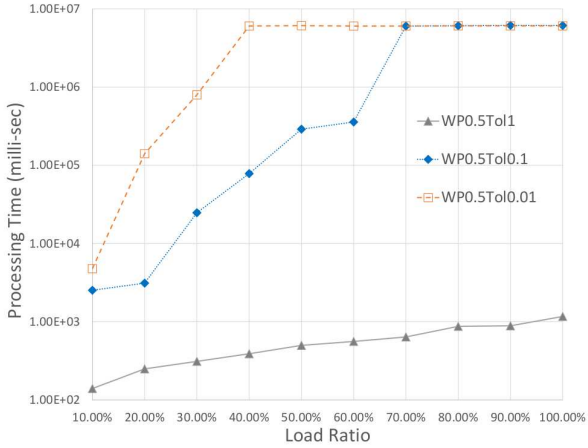

Fig. 13. Complexity of Solving Algorithm versus Load and Several Optimality Tolerance Gaps.

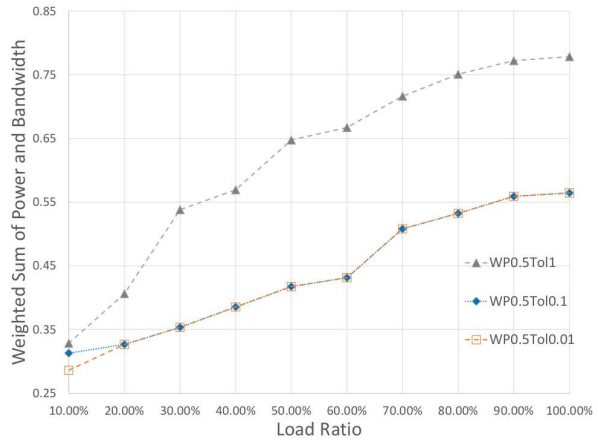

Fig. 14. Objective Function Evaluation for Several Optimality Tolerance Gaps, Associated with Fig. 13.

Fig. 13 evaluates the complexity of constraint programming solver as the load ratio increases for different optimality tolerance gap (noted as 'Tol'={0.01,0.1,1}[19]). The optimality tolerance gap governs how closely the optimization solver must approach the theoretically optimal solution, i.e., the lower value of 'Tol' the closest to the theoretical limit we get. This simulation is conducted on Intel(R) Xeon(R) CPU E5-1650 v3 with clock speed 3.5GHz, 64 GB RAM, and Windows operating system. Fig. 14 shows the weighted sum of power and bandwidth (objective function) that is associated with the complexity analysis of Fig. 13, for a system of 43 RUs, and four RSs, and number of users increases from 17 to 172 (low to high load ratio). The time complexity increases from 5e3 to 9e6 msec if 'Tol'=0.01 as the load increases. By increasing the tolerance gap to 'Tol'=1 the processing time drops to the range of 140 msec to 1170 msec, which is much lower than that of 0.01 gap case. However, this comes at the expense of degrading the performance in the range of 14.69% to 37.5%, as described in the associated Figure 14. Finally, we note that increasing the load more than 40% (for 'Tol'=0.01) does not change the processing time. This is because the optimal solution of a higher load is similar, i.e., optimal allocation of processing functions will not change under similar settings by changing the load. This conclusion can be inferred from Fig. 6.

The applicability of such an optimization tool depends on the problem nature. Our framework addresses a design problem that outputs the optimal processing allocation for future radio networks given the available midhaul bandwidth, the capacity of existing digital units, and the number of remote sites as input to the problem. Since this is a design problem, which is necessary at the early stages of deploying such a network, the needed solution time is not on the scales of milliseconds. Hence, this tool is applicable to our framework, given the reported processing time in Fig. 13.

V. Conclusion

In this paper, we considered a hybrid cloud radio access network architecture with dual-site processing with three hierarchical layers. In the proposed system, baseband processing chain can be split, and virtualized functions can be provisioned at both central and remote sites. We modeled the functional split in H-CRAN, as a sequence of cell processing functions and user processing functions. We proposed an optimization framework to jointly minimize the power consumption and transport bandwidth consumption, by developing a constraint programming model. Numerical results showed that when power consumption is more valued, as more transport bandwidth capacity is available, more functions are placed at the CS to save power. Also, the interplay of power and bandwidth consumptions is drastic, and there exists a balanced point for joint minimization of them, where the weighting factor is about 0.5-0.65. We found that when valuing the power minimization the most, under high system load, it is recommended to use a higher number of RSs. We reported that the system power consumption at low load is lower than the high load by about 20.5 to 51 percentile. System bandwidth consumption at low load is lower than high load by about 9.3 to 81.6 percentile. In the future, we want to consider a dynamic scenario where users arrive/depart at/from the network, with different service requirements, e.g., latency. Also, users' services requirements impact on such a framework will be considered. This problem is similar to that of the bin packing one, which is NP-hard. Hence the optimal solution to this problem with the CP will not be scalable for large problem sizes. Therefore in a real-time scenario, more time-efficient algorithms need to be developed as a future work. We are currently considering the impact of changing radio access parameters, e.g., modulation

[19]Note that we did not report lower gaps than 0.01 because there is no performance improvement found in lower gaps.

index and physical resource block, etc., on the overall energy consumption and latency per required service.

REFERENCES

[1] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5G be?" *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, June 2014.

[2] 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, Firstquarter 2015.

[3] M. Kamel, W. Hamouda, and A. Youssef, "Ultra-Dense Networks: A Survey," *IEEE Communications Surveys Tutorials*, vol. PP, no. 99, pp. 1–1, 2016.

[4] X. Wang, S. Thota, M. Tornatore, H. S. Chung, H. H. Lee, S. Park, and B. Mukherjee, "Energy-Efficient Virtual Base Station Formation in Optical-Access-Enabled Cloud-RAN," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1130–1139, May 2016.

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

[6] 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.

[7] A. Maeder, M. Lalam, A. D. Domenico, E. Pateromichelakis, D. Wbben, 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*, June 2014, pp. 1–5.

[8] J. Liu, S. Zhou, J. Gong, Z. Niu, and S. Xu, "Graph-based framework for flexible baseband function splitting and placement in C-RAN," in *2015 IEEE International Conference on Communications (ICC)*, June 2015, pp. 1958–1963.

[9] 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.

[10] J. Liu, S. Xu, S. Zhou, and Z. Niu, "Redesigning fronthaul for next-generation networks: beyond baseband samples and point-to-point links," *IEEE Wireless Communications*, vol. 22, no. 5, pp. 90–97, October 2015.

[11] 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.

[12] M. Vincenzi, A. Antonopoulos, E. Kartsakli, J. Vardakas, L. Alonso, and C. Verikoukis, "Cooperation incentives for multi-operator c-ran energy efficient sharing," in *2017 IEEE International Conference on Communications (ICC)*, May 2017, pp. 1–6.

[13] G. Tseliou, F. Adelantado, and C. Verikoukis, "Scalable ran virtualization in multitenant lte-a heterogeneous networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 8, pp. 6651–6664, Aug 2016.

[14] H. Beyranvand, M. Lvesque, M. Maier, J. A. Salehi, C. Verikoukis, and D. Tipper, "Toward 5G: FiWi Enhanced LTE-A HetNets With Reliable Low-Latency Fiber Backhaul Sharing and WiFi Offloading," *IEEE/ACM Transactions on Networking*, vol. 25, no. 2, pp. 690–707, April 2017.

[15] A. Bousia, E. Kartsakli, A. Antonopoulos, L. Alonso, and C. Verikoukis, "Multiobjective auction-based switching-off scheme in heterogeneous networks: To bid or not to bid?" *IEEE Transactions on Vehicular Technology*, vol. 65, no. 11, pp. 9168–9180, Nov 2016.

[16] 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.

[17] 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.

[18] "40-Gigabit-capable passive optical networks (NG-PON2)," ITU-T G.989 series of Recommendations:, ITU-T, March 2013.

[19] 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.

[20] "Further Study on Critical C-RAN Technologies." Release, Next generation mobile network alliance, Mar. 2015.

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

[22] U. Dtsch, M. Doll, H. P. Mayer, F. Schaich, J. Segel, and P. Sehier, "Quantitative analysis of split base station processing and determination of advantageous architectures for lte," *Bell Labs Technical Journal*, vol. 18, no. 1, pp. 105–128, June 2013.

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

[24] "Use constraint programming to compute optimized schedules and solve other hard optimization problems."

[25] X. Wang, "UC Davis Technical Report. [online]:http://networks.cs.ucdavis.edu/xinbo/appendix-a-techno-economic-study-to-design-low-cost-edge-cloud-radio-access-network.pdf," 2016.