Optimizing functional split of baseband processing on TWDM-PON based fronthaul network

Go Hasegawa
Research Institute of Electrical Communication
Tohoku University, Sendai, JAPAN
hasegawa@riec.tohoku.ac.jp

Masayuki Murata
Graduate School of Information Science and Technology
Osaka University, Osaka, JAPAN
murata@ist.osaka-u.ac.jp

Yoshihiro Nakahira and Masayuki Kashima
Oki Electronics Industry Co., Ltd.
Saitama, JAPAN
{nakahira523, kashima567}@oki.com

Shingo Ata
Graduate School of Engineering
Osaka City University, Osaka, JAPAN
ata@info.eng.osaka-cu.ac.jp

***Abstract*—One of the major shortcomings of Centralized Radio Access Networks (C-RAN) is that the large capacity is required for fronthaul network between Remote Radio Heads (RRHs) and central office with baseband unit (BBU) pool. Possible solutions are to introduce lower-cost networking technology for fronthaul network, such as Time and Wavelength Division Multiplexing Passive Optical Network (TWDM-PON), and to introduce functional split, that moves some baseband processing functions to cell site to decrease the utilization of the fronthaul network. In this paper, we construct the mathematical model for selecting function split options of baseband processing to minimize the power consumption of TWDM-PON based fronthaul network. In detail, we formulate the optimization problem for minimizing the total power consumption of fronthaul network in terms of the capacity of TWDM-PON, the number of RRHs in each cell site, server resources, latency constraints, the amount of traffic from each RRH, physical/virtual server power consumption characteristics. Numerical examples are shown for confirming the correctness of the proposed model and for presenting the effect of resource enhancement methods on the capacity and energy efficiency of the system.**

***Index Terms*—5G, fronthaul network, TWDM-PON, functional split, baseband processing**

I. Introduction

Centralized/Cloud-based Radio Access Network (C-RAN) is often utilized in Long Term Evolution (LTE) and fifth-generation (5G) mobile cellular networks, where baseband processing is executed at a central office with Baseband Unit (BBU), not at base stations as in the traditional Distributed RAN (D-RAN). Its objectives include the decrease of the cost of base stations and the introduction of coordinated controls of multiple base stations for enhancing radio network performance [1], [2]. Furthermore, by introducing cloud technologies to the BBU, meaning that the baseband processing is executed by software on general-purpose servers, we can expect high efficiency in resource utilization and power consumption because adaptive utilization of virtual/physical servers is achieved according to the cell configuration and the traffic amount.

978-1-7281-5127-4/20/$31.00 ©2020 IEEE

In C-RAN, Remote Radio Heads (RRHs) and BBU pool are interconnected via a fronthaul network, and physical radio data is transmitted from RRHs to BBU pool by specialized interface such as Common Public Radio Interface (CPRI). Because of the characteristics of CPRI, the required bandwidth is quite large (e.g. 2.46 Gbps for each 20MHz channel bandwidth), and is independent on the amount of user traffic. When massive multi-input multi-output (MIMO) technologies and ultra-dense cell configuration are introduced in 5G networks, this problem becomes more serious. To decrease the large cost of fronthaul networks, ethernet-based fronthaul networks such as Next Generation Fronthaul Interface (NGFI) [3] and Time and Wavelength Division Multiplexing Passive Optical Network (TWDM-PON) based fronthaul network [4] are possible solutions.

On the other hand, functional split is now paid much attention, where a part of baseband processing functions are executed at cell sites, the processed data is transmitted via the fronthaul network, and the remaining processing functions are conducted at the central office [5], [6]. By executing lower-layer processing at RRH, the required bandwidth on fronthaul network is significantly decreased, that can compress the cost of fronthaul network for realizing 5G networks. Furthermore, by introducing software-based processing at the cell site, on-demand and adaptive configuration of functional split can be realized. However, by introducing server resources at the cell sites, the power consumption and monetary cost of the system may increase. The functional split options, meaning that what part of baseband processing functions are executed at each of the cell site and central office, should be determined according to various factors such as capacity and power consumption characteristics of servers and fronthaul network, application demands, and traffic amount. Especially, when TWDM-PON is utilized for the fronthaul network, various characteristics of TWDM-PON, such as the shared-bandwidth nature, should be taken into account. However, to the best of our knowledge, there is no previous work on selecting functional split options for baseband processing on TWDM-PON based fronthaul

network.

In this paper, we give a mathematical model of selecting functional split options for baseband processing in TWDM-PON based fronthaul network. In detail, we formulate the optimization problem in Integer Linear Program (ILP) form to minimize the power consumption of the whole system by selecting functional split options, in terms of the capacity and latency of TWDM-PON network, the distribution of RRHs, traffic amount from each RRH, performance and power consumption characteristics of servers at cell sites and central office, as well as the applications' requirement on end-to-end latency. Through numerical examples we exhibit that different resource enhancement methods for performance improvement gives the different effect on the energy efficiency, that can be analyzed by our mathematical model.

The rest of this paper is organized as follows. Section II summarizes the related work. In Section III we give the network and system model, functional split model, and power consumption model. Then we formulate the optimization problem for selecting functional split options in Section IV. Section V exhibits the numerical examples of the analysis model and gives some discussions. Finally, the conclusions of this paper and some future work are presented in Section VI.

II. Related work

There are some previous works on the function split of baseband processing in LTE/5G networks. In [7], the authors discuss the effect of the functional split on the load of the fronthaul network, considering the cell configuration and multiplexing effect. They also focus on the effect of packetization of the fronthaul network. However, they did not consider the queueing effect of the packet-based network. In this work, we assume the TWDM-PON based fronthaul network, where the network bandwidth is fixedly allocated by Dynamic Bandwidth Allocation (DBA) algorithm in Passive Optical Network (PON).

In [8], the authors consider the functional split problem as a Virtual Network Embedding (VNE) problem, and formulate the problem as an ILP to obtain the virtual network configuration with optimal functional split options according to the requirements from Mobile Virtual Network Operators (MVNOs). However, the model for server performance is quite simple, and detailed networking technologies and power consumption are not considered for fronthaul network. In the present paper, we considered the detailed configurations and power consumption of TWDM-PON, and server performance is determined from the existing experimental results of baseband processing.

The authors in [9] formulate the optimization problem for minimizing the power consumption of the mobile network system, considering the effect of the offloading baseband processing functions and user tasks to fog and cloud servers. To obtain numerical evaluation results, the authors exploit the numerical data on the overhead of baseband processing presented in [10]. However, the detailed networking technologies are not considered for fronthaul and backhaul networks.

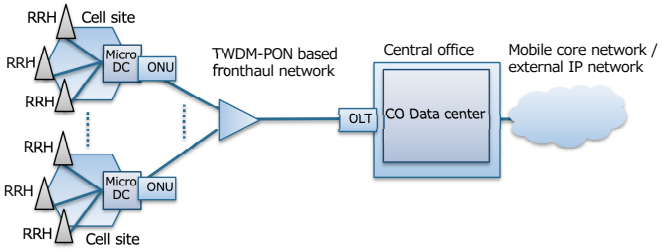

Fig. 1. Network and system model

Also, the power consumption model they utilized is too simple to capture the actual characteristics of power consumption of servers and network equipment. Our work exploits the power consumption model for servers found in [11]. Also, we formulate the power consumption of TWDM-PON network based on its detailed physical configuration.

In [12], the authors consider TWDM-PON based fronthaul network, and formulate the optimization problem to minimize the power consumption by selecting the location at which baseband processing is executed, from fog servers at cell site and cloud servers at central office. However, the model does not consider the network topology and the configuration of cell site. Also, they ignore the network traffic on the fronthaul network after baseband processing. Our model is similar to that in [12], but we construct more realistic model by considering various factors such as functional split options, the traffic amount on the fronthaul network, the network topology, and cell site configuration. Also, for obtaining numerical evaluation results, we borrowed the experimental results in [13] and [14] which conducted the experiments of functional split in LTE networks. Also, in [15], the experimental results on functional split with virtual machines are presented to determine the limitation of the transmission latency of fronthaul network. However, these works do not discuss the optimal functional split options.

III. Analysis model

A. Network and system model

Figure 1 depicts the network and system model used in this work. One or multiple cell sites, each of which accommodates one or several RRHs, are connected to a central office via a TWDM-PON based fronthaul network. The network traffic generated at each RRH is transmitted to the central office via Optical Network Unit (ONU) at the entrance of the cell site and Optical Line Terminal (OLT) at the entrance of the central office. The TWDM-PON based fronthaul network supports multiple wavelength and each ONU can choose the wavelength for each RRH traffic. The cell sites and central office has a cluster of physical servers, which is called as micro data center ("Micro DC" in the figure) and CO data center, respectively.

For the network traffic generated at each RRH, the baseband processing is executed both of or either one of micro data cen-

ter or CO data center, and resulting IP packets are forwarded toward the mobile core network or external IP network.

B. Functional split options for baseband processing

In this work, the functional split options are defined based on [13], [14]. Figure 2 depicts four options for functional split (Split 1–4) for uplink network traffic from a RRH to mobile core network. Each option has the functions executed at the micro data center at the cell site and those executed at the CO data center at the central office. Split 1 corresponds to C-RAN configuration, where almost all functions are executed at the central office, whereas Split 4 is D-RAN configuration where all functions are executed at the cell site. Split 2 and Split 3 are midway options, each of them has different portion of functions executed at the cell site and central office.

In each split option, the load on servers and processing latency at micro data center and CO data center are different. Also, the required bandwidth of the fronthaul network is dependent on the portion of functions executed at the micro data center. Table I summarizes the CPU load, processing latency, and required bandwidth for each split option. We determined the processing latencies and CPU utilizations based on the experimental results in [13], [14], respectively. In detail, "Low performance server" and "High performance server" correspond to the server used in the experiments in [14] and [13], respectively, and we utilized the experimental results in these works and converted them appropriately based on CPU benchmark scores in [16]. Note that when the CPU utilization is larger than 100%, it means that the functions require multiple CPU cores to be executed. The required network bandwidths are the calculation results in [14].

From Table I, we can observe that from Split 1 through Split 4, the CPU overhead at the micro data center increases, while that at the CO data center decreases. The processing latency changes according to the CPU overhead. These changes also affect the power consumption at both data centers. Furthermore, the required bandwidth of fronthaul network decreases from Split 1 through Split 4, that affect the load on the fronthaul network. Therefore, for optimizing the system performance, we should select a functional split option for each RRH traffic carefully according to various factors such as the capacity and the power consumption characteristics of servers and fronthaul networks.

C. Power consumption model

Servers at micro data centers and CO data center consume electrical power when executing baseband processing functions. In this work we exploit the power consumption model in [11] as shown in the following equation, where $P(x)$ is the power consumption of a virtual server when the load is x.

$$P(x) = \begin{cases} 0 & x = 0 \\ \frac{M-I}{G}x + I & \text{otherwise}, \end{cases} \quad (1)$$

where G is the maximum performance of a virtual server, M is the power consumption at the maximum performance, I is the power consumption when the server has no load. In this model, the power consumption increases linearly as the server load increases. When we can stop a virtual server with no load, its power consumption is zero.

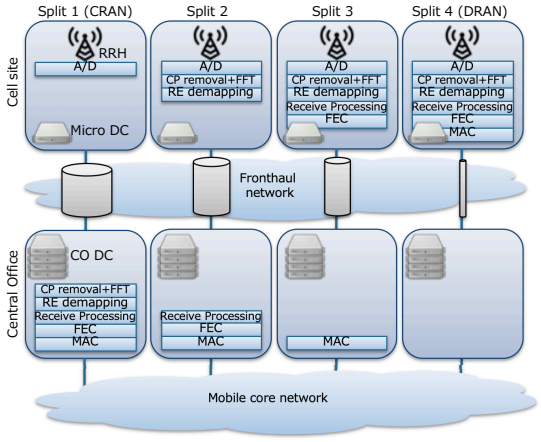

Fig. 2. Functional split options for baseband processing [14]

The power consumption of TWDM-PON based fronthaul network is determined the number of wavelengths activated for accommodating RRH traffic. It means that the number of used wavelengths should be minimized to decrease the power consumption of the network.

Note that our analysis model can accept different power comsumption model, since it is interpreted into pre-defined numerical paremeters for optimication problem.

IV. Optimization problem formulation

In this section, we formulate the selecting problem of functional split options as an ILP that minimizes the total power consumption required for processing the network traffic generated at RRHs, based on the network and system model in Section III.

A. Variable definitions

The variables in the model are defined as follows. R is the number of RRHs in the network, each of which is denoted as RRH_i ($1 \leq i \leq R$). Cell sites and a central office are called as nodes. The total number of nodes is N, where the central office is node 1 and the cell sites corresponds to node 2 through N. a_n^i ($1 \leq i \leq R$, $2 \leq n \leq N$) explains the network topology, where $a_n^i = 1$ when RRH_i is accommodated in node n and $a_n^i = 0$ otherwise.

The number of wavelengths of TWDN-PON is W, and the network bandwidth of the wavelength w ($1 \leq w \leq W$) is denoted by B_w. The number of split options is K and each split is denoted by Split k ($1 \leq k \leq K$). $B_{n,k,w}^i$ means the required network bandwidth when the traffic from RRH_i from node n is processed by Split k and transmitted by the wavelength w. $X_{n,k}$ ($2 \leq n \leq N$, $1 \leq k \leq K$) means the power consumption at the micro data center (node n) when the traffic from RRH is processed by Split k. $X_{1,k}$ ($1 \leq k \leq K$) means the power consumption at the data center (node 1) when the

TABLE I
RESOURCE CONSUMPTION AND PROCESSING LATENCY OF SPLIT OPTIONS

	Low performance server				High performance server			
Split option	Split 1	Split 2	Split 3	Split 4	Split 1	Split 2	Split 3	Split 4
CPU utilization at micro data center [%]	0	32.2	126.7	139.5	0	12.4	48.7	53.7
Processing latency at micro data center [μs]	0	111	1,772	1,972	0	43	682	758
CPU utilization at CO data center [%]	139.5	107.3	12.8	0	53.7	41.3	4.9	0
Processing latency at CO data center [μs]	1,972	1,861	200	0	758	716	77	0
Traffic rate [Gbps]	2.46	0.72	0.054	0.054	2.46	0.72	0.054	0.054

traffic from RRH is processed by Split k. $D_{n,k}$ ($2\leq n \leq N$, $1\leq k \leq K$) and $D_{1,k}$ ($1\leq k \leq K$) represents the processing latency at the micro data center (node n) and CO data center (node 1), respectively, when the traffic from RRH is processed by Split k.

C_n ($1\leq n \leq N$) is the power consumption of the physical server of node n. L_w ($1\leq w \leq W$) means the power consumption when wavelength w of TWDM-PON is activated for accommodating one or more RRH traffic. τ_n ($2\leq n \leq N$) is the propagation delay of TWDM-PON between node 1 and node n. P_n ($1\leq n \leq N$) is the number of CPU cores at the node n. When the number of cores is one for each server, the number of CPU cores corresponds to the number of servers.

B. Optimization problem definition

The optimization problem to minimize the total power consumption of the system is formulated as follows.

$$\text{Minimize :}$$
$$\sum_{n=2}^{N}\sum_{i=1}^{R}\sum_{k=1}^{K}\sum_{w=1}^{W} y^i_{n,k,w}(X_{n,k}+X_{1,k}) + \sum_{n=1}^{N} x_n C_n + \sum_{w=1}^{W} l_w L_w \quad (2)$$
$$\text{Subject to :}$$
$$y^i_{k,n,w}, x_n, l_w \in \{0,1\} \quad (3)$$
$$y^i_{k,n,w} \leq a^i_n \ \forall i,n,k,w \quad (4)$$
$$\sum_{n=2}^{N}\sum_{k=1}^{K}\sum_{w=1}^{W} y^i_{n,k,w} = 1 \ \forall i \quad (5)$$
$$\sum_{n=2}^{N}\sum_{i=1}^{R}\sum_{k=1}^{K} y^i_{n,k,w} B^i_{n,k,w} \leq B_w \ \forall w \quad (6)$$
$$y^i_{n,k,w}\{D_{n,k}+\tau_n+D_{1,k}\} \leq \Delta_i \ \forall i,n,k,w \quad (7)$$
$$\sum_{k=1}^{K}\sum_{i=1}^{R}\sum_{w=1}^{W} y^i_{n,k,w}\rho_k < 100\cdot P_n \ \forall n, \quad (8)$$

where decision variables are $y^i_{k,n,w}$, x_n, and l_w. $y^i_{k,n,w}$ is 1 when the traffic from RRH_i at node n is processed by Split k and transmitted by the wavelength w, and 0 otherwise. x_n is 1 when the physical server at node n is activated, and 0 otherwise. l_w is 1 when the wavelength w is used for accommodating RRH traffic, and 0 otherwise. Note that x_n and l_w are determined from $y^i_{k,n,w}$ as follows.

$$x_n = \begin{cases} 1 & \text{if } \sum_{i=1}^{R}\sum_{k=1}^{K}\sum_{w=1}^{W} y^i_{n,k,w} \geq 1 \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

$$l_w = \begin{cases} 1 & \text{if } \sum_{n=2}^{N}\sum_{i=1}^{R}\sum_{k=1}^{K} y^i_{n,k,w} \geq 1 \\ 0 & \text{otherwise.} \end{cases} \quad (10)$$

Equation (2) consists of the power consumption of virtual servers at micro data centers and CO data center, the power consumption of physical servers, and the power consumption of TWDM-PON fronthaul network. Equation (3) means that the decision variables must be binary. Equation (4) represents the network topology constraints, meaning that the traffic from a RRH must be processed at the micro data center at the cell site where the RRH is accommodated. Equation (5) is the constraint so that all RRH traffic must be processed one and only one micro data center and the central office. Equation (6) means that the amount of traffic accommodated by each wavelength of TWDN-PON must be equal to or smaller than the capacity of the wavelength. Equation (7) represents the end-to-end latency constraint, where the sum of processing latencies at the micro data center and CO data center, and the propagation delay of TWDM-PON must be equal to or smaller than the upper limit determined for each RRH traffic. Finally, Equation (8) means that the total load on each server must be equal to or smaller than the server capacity.

V. NUMERICAL EVALUATIONS

In this section, the numerical examples of the optimization problem in Section IV are presented to confirm its effectiveness to derive the optimal configurations of functional split options to minimize the power consumption. We also consider the scenarios of resource enhancement for performance improvement and their effect on the energy efficiency.

The optimization problem is solved by IBM® ILOG® CPLEX® Interactive Optimizer 12.10.0.0 [17] on MacBook Pro® 2019 (2.4 GHz and 8 core of Intel Core i9 CPU and 64 GB of 2,667 MHz DDR4 memory). The calculation time for each optimication problem is less than 200 msec. When there are multiple solutions that give the minimal power

consumption, we utilize one of them to present the evaluation results.

A. Parameter settings and evaluation scenarios

We utilize the network model in Figure 1, and set the number of nodes (N) to 2, meaning that there is only one cell site in the network. Note that the cases of $N > 2$ give the similar results to what follows. The link bandwidth of each TWDM-PON wavelength w (B_w) is set to 10 [Gbps]. The traffic amount from each RRH is set to 0.054 [Gbps] as in Table I.

The power consumption for activating each TWDM-PON wavelength w (L_w) is set to 20 [W]. The number of split options (K) is four as in Table I. $X_{n,k}$, $D_{n,k}$, and $B^i_{n,k,w}$ are also configured as in Table I. Baseband processing at micro data centers and CO data center is executed on virtual machines activated on the physical servers. The physical server at the central office is always activated, and its power consumption (C_1) is 200 [W]. The physical servers at the micro data centers are activated only when one or more virtual servers are activated.

The power consumption for executing baseband processing functions are determined by Equation (1) in Subsection III-C, where x means the CPU utilization [%] and its upper limit (G) is determined by the product of the number of CPU cores at node n (P_n) and 100. The power consumption at the maximum performance (M) is determined from the Thermal Design Power (TDP) of CPU found in [16], where the power consumption is equal to TDP when the loads of all CPU cores are 100 [%]. I is set to a half of M.

For the baseline scenario (Scenario 1), the following parameters are utilized: W = 3, P_1 = 30, P_2 = 3, D_n = 2,000 [μsec] ($2 \leq n$), τ_n = 10 [μsec] ($2 \leq n$). In this scenario, the micro data centers have low performance servers, and CO data center has a high performance server.

Furthermore, we consider the multiple scenarios where the system resource is enhanced for increasing the number of accommodated RRHs. In Scenario 2, the server at the micro data centers are changed to high performance servers. In Scenario 3, we increase the number of CPU cores at the micro data center (P_2) from 3 to 8. In Scenario 4, we increase the number of wavelengths (W) from 3 to 6. These parameters are determined so that the number of accommodated RRHs becomes roughly equal. We compare the power consumption of each scenario and discuss the effect of the resource enhancing methods on the system performance and the energy efficiency.

B. Evaluation results and discussions

In Figure 3, we present the evaluation results of Scenario 1. The upper graph depicts the usage of function split options and TWDM-PON wavelengths, as a function of the number of RRHs to be accommodated. Each colored box in the graph means the utilization of a split option for processing a RRH traffic. For example, when the number of RRHs is five, two RRH traffic are processed by Split 1, one by Split 2, one by Split 3, and one by Split 4. The number of vertical bars means the number of wavelengths used to accommodate all RRH traffic. When the number of RRHs is less than 11, the number of used wavelengths is one. The number of wavelengths is two when the number of RRHs is ranged from 11 to 14. When the number of RRHs is ranged from 15 to 18, the number of used wavelengths is three. Finally, then the number of RRHs is larger than 18, the CPLEX solver has no solution, that means that the system cannot accommodated 19 or more RRHs. Therefore, the baseline scenario (Scenario 1), the upper limit of the number of accommodated RRH traffic is 18. The lower graph represents the change in the total power consumption.

From the figures, we can observe that when the number of RRHs is small, Split 3 and Split 4 are often used. This is because the micro data center has a lower performance server, that gives smaller power consumption than the high performance server at CO data center. So, to decrease the total power consumption, larger portion of baseband processing functions are executed at the micro data center. On the other hand, when the number of RRHs increases, most of traffic are processed by Split 1 and Split 2, that results in the increase of the number of used wavelengths. This is because the lack of server capacity at micro data centers.

We can also find that the power consumption increases almost linearly as the number of RRHs increases. Additional increases are found when the number of used wavelengths increases.

Figures 4–6 present the evaluation results for Scenarios 2–4, respectively. For comparison purposes, we plot the power consumption of Scenario 1 in the lower graph of each figure. The upper limits of the number of accommodated RRH traffic in Scenarios 2–4 are 29, 29, and 30, respectively.

In Scenario 2 (Figure 4), we can confirm that Split 3 is often used compared with Scenario 1. This is because the server at the micro data center is enhanced to high performance server and more baseband processing functions can be executed. This also contributes the reduction of the number of used wavelengths. For example, when the number of RRHs is 15, only one wavelength is used in Scenario 2, while three wavelengths are used in Scenario 1. The total power consumption increases as compared with Scenario 1, that is caused by increased power consumption of the server at the micro data center.

In Scenario 3 (Figure 5), we can confirm that Split 3 and Split 4 are often used compared with Scenario 1. This is also because of the resource enhancing at the micro data center. We also note that the power consumption is smaller than Scenario 1 when the number of RRHs is ranged from 11 to 18. This is because of the reduction of the number of used wavelengths.

While we can observe the almost the same effect on the upper limit of the number of the accommodated RRHs and the number of used wavelengths in Scenarios 2 and 3, the total power consumption of Scenario 3 is significantly small (approx. 20% reduction) compared with that of Scenario 2. This means the importance of the selection of resource en-

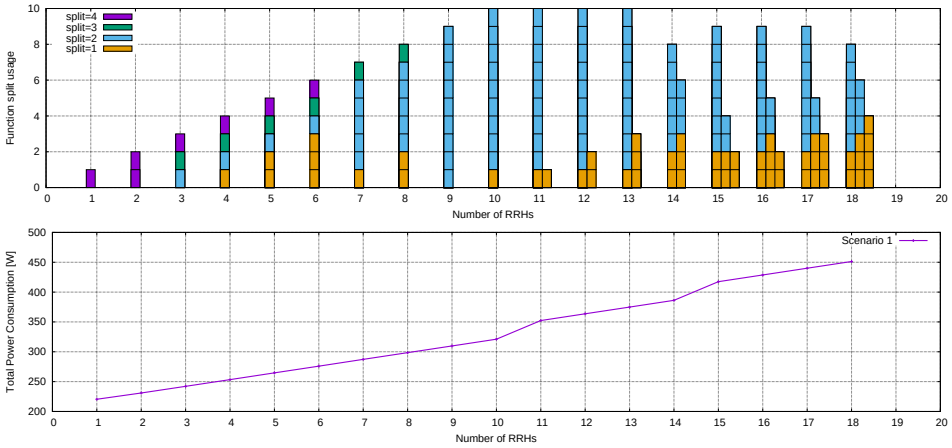

Fig. 3. Evaluation results of Scenario 1 (Upper: Function split usage, Lower: Total power consumption)

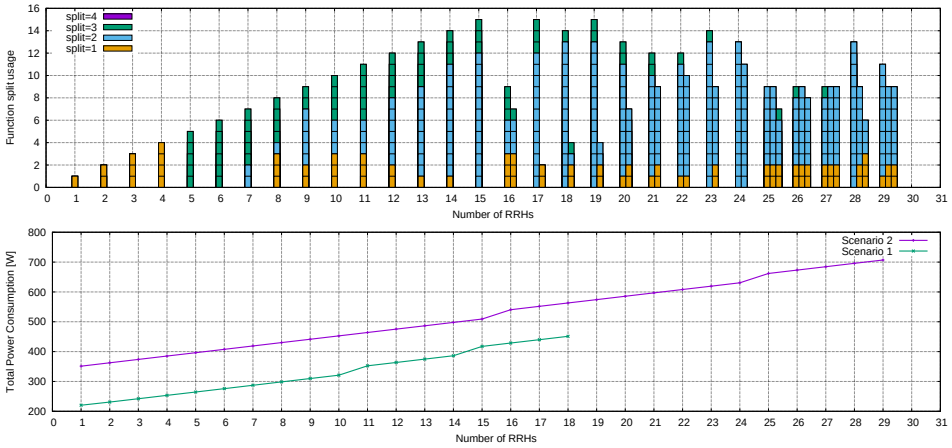

Fig. 4. Evaluation results of Scenario 2 (Upper: Function split usage, Lower: Total power consumption)

hancing method when we care about the power consumption of the system. Such optimization can be easily achieved by solving the optimization problem defined in this paper.

In Scenario 4 (Figure 6), the selected split options and the total power consumption remain unchanged from the results in Scenario 1, since the server performance is identical. However, since the number of wavelengths increases, the number of RRHs that can be accommodated is enhanced. Furthermore, when comparing the power consumption with Scenarios 2 and 3, Scenario 4 exhibits middle performance among the three scenarios. We also note that the power consumption variation in the three scenarios is around 140 [W], that is roughly 25 [%] of the power consumption of Scenario 3. This result again strengthens the importance of selecting resource enhancing method on the energy efficiency of the system.

VI. Conclusions and Future Work

In this paper, we gave the mathematical model for selecting functional split options of baseband processing for optimizing the system performance with fronthaul network based on TWDM-PON. In detail, we formulated the selecting problem of functional split options as an ILP to minimize the power consumption, with constraints on the network and server capacities, end-to-end latency, and the network topology, considering the power consumption characteristics. Through numerical examples, we discussed the effect of resource enhancing methods for performance improvement on the energy efficiency. One of the results is that adding CPU cores at micro data centers is effective to increase the number of RRHs to be accommodated while the total power consumption kept small.

In this work, for obtaining numerical evaluation results we utilized the experimental results of function split in LTE networks in [13], [14]. Since our model is independent on the mobile network generation, the numerical evaluation for 5G networks can be conducted when we have experimental results of function split in 5G network. This is one important future work.

We also plan to investigate the scalability of our optimication problem in terms of the number of entities and traffic demends of the system, that is important for short-term adaptation to the dynamically-changing network environment. It may be affected by the implementation difficulties of functional split technologies and the overhead in gathering information

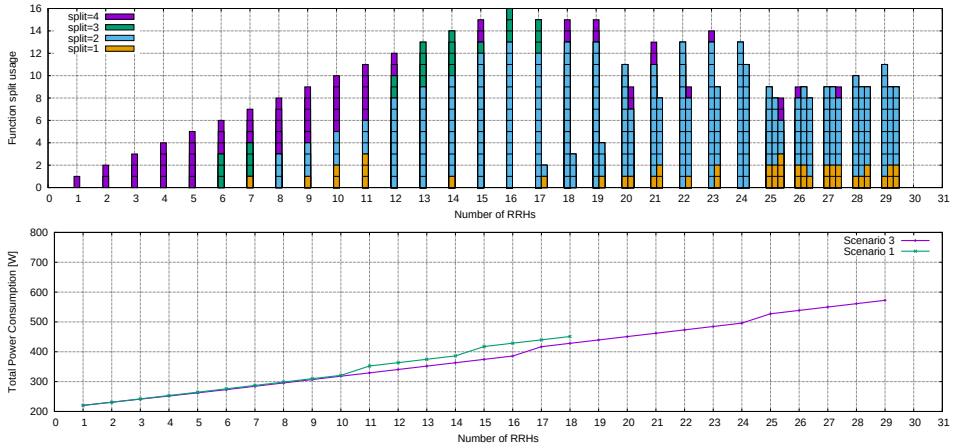

Fig. 5. Evaluation results of Scenario 3 (Upper: Function split usage, Lower: Total power consumption)

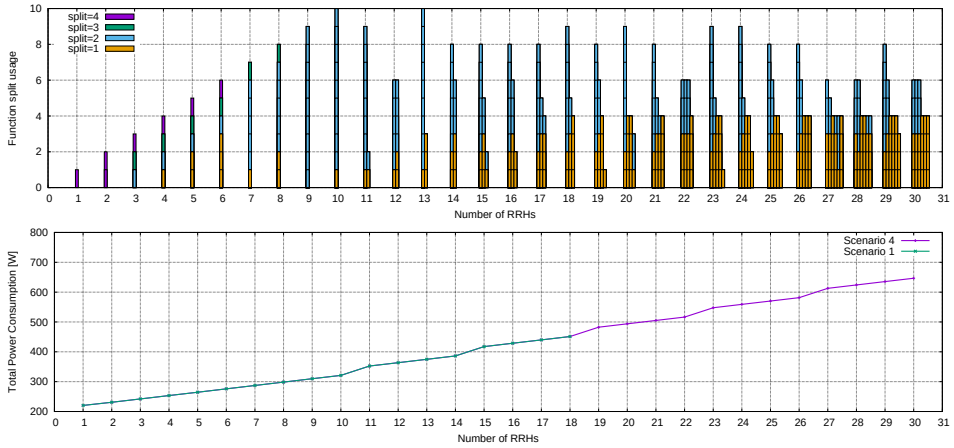

Fig. 6. Evaluation results of Scenario 4 (Upper: Function split usage, Lower: Total power consumption)

for solving the optimization problem.

REFERENCES

[1] J. Wu, Z. Zhang, Y. Hong, and Y. Wen, "Cloud radio access network (C-RAN): a primer," *IEEE Network*, vol. 29, pp. 35–41, Jan. 2015.

[2] M. Peng, Y. Sun, X. Li, Z. Mao, and C. Wang, "Recent advances in cloud radio access networks: System architectures, key techniques, and open issues," *IEEE Communications Surveys & Tutorials*, vol. 18, pp. 2282–2308, Mar. 2016.

[3] IEEE Standards Association, "P1914.1 - IEEE draft standard for packet-based fronthaul transport networks," *available from https://standards.ieee.org/project/1914_1.html*.

[4] T. Pfeiffer, "Next generation mobile fronthaul and midhaul architectures," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 7, pp. B38–B45, Nov. 2015.

[5] L. M. P. Larsen, A. Checko, and H. L. Christiansen, "A survey of the functional splits proposed for 5G mobile crosshaul networks," *IEEE Communications Surveys & Tutorials*, vol. 21, pp. 146–172, Oct. 2018.

[6] M. A. Habibi, M. Nasimi, B. Han, and H. D. Schotten, "A comprehensive survey of RAN architectures toward 5G mobile communication system," *IEEE Access*, vol. 7, pp. 70371–70421, May 2019.

[7] C.-Y. Chang, R. Schiavi, N. Nikaein, T. Spyropoulos, and C. Bonnet, "Impact of packetization and functional split on C-RAN fronthaul performance," in *Proceedings of ICC 2016*, May 2016.

[8] D. Harutyunyan and R. Riggio, "Flexible functional split in 5G networks," in *Proceedings of CNSM 2017*, Nov. 2017.

[9] Z. Cheng, Y. Tang, and H. Wu, "Joint task offloading and flexible functional split in 5G radio access network," in *Proceedings of ICOIN 2019*, Jan. 2019.

[10] C. Desset, B. Debaillie, V. Giannini, A. Fehske, G. Auer, H. Holtkamp, W. Wajda, D. Sabella, F. Richter, M. J. Gonzalez, H. Klessig, I. Gódor, M. Olsson, M. A. Imran, A. Ambrosy, and O. Blume, "Flexible power modeling of LTE base stations," in *Proceedings of IEEE WCNC 2012*, Apr. 2012.

[11] P. Mahadevan, P. Sharma, S. Banerjee, and P. Ranganathan, "A power benchmarking framework for network devices," in *Proceedings of Networking 2009*, May 2009.

[12] R. I. Tinini, L. C. M. Reis, D. M. Batista, G. B. Figueiredo, M. Tornatore, and B. Mukherjee, "Optimal placement of virtualized BBU processing in hybrid cloud-fog RAN over TWDM-PON," in *Proceedings of GLOBECOM 2017*, Dec. 2017.

[13] N. Nikaein, "Processing radio access network functions in the cloud: Critical issues and modeling," in *Proceedings of MCS 2015*, Sept. 2015.

[14] M. Kist, J. A. Wickboldt, L. Z. Granville, J. Rochol, L. A. DaSilva, and C. B. Both, "Flexible fine-grained baseband processing with network functions virtualization: Benefits and impacts," *Computer Networks*, vol. 151, pp. 158–165, Mar. 2019.

[15] F. Giannone, H. Gupta, K. Kondepu, D. Manicone, A. Franklin, P. Castoldi, and L. Valcarenghi, "Impact of RAN virtualization on fronthaul latency budget: An experimental evaluation," in *Proceedings of GLOBECOM 2017*, Dec. 2017.

[16] IBM, "CPU benchmarks," *available from https://www.cpubenchmark.net/*.

[17] IBM, "IBM ILOG CPLEX optimization studio," *available from https://www.ibm.com/jp-ja/products/ilog-cplex-optimization-studio*.