

System and application performance of function placement strategies for virtualized mobile fronthaul/backhaul networks

Go Hasegawa

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

Rina Yamasaki and Masayuki Murata

Graduate School of Information Science and Technology
Osaka University, Osaka, JAPAN
{r-yamask, murata}@ist.osaka-u.ac.jp

Abstract—In this paper, we focused on the function placement problem of mobile cellular networks on the premise of integrated control of the fronthaul and the backhaul network. We construct the analytical model of the problem to derive the end-to-end delay and packet loss rate for application traffic, and the power consumption of the whole system. Through numerical evaluations, we present that by properly placing the network functions of the fronthaul and the backhaul networks, it is possible to reduce the end-to-end delay time of application traffic significantly, while the power consumption and packet loss rate remain almost unchanged.

Index Terms—Mobile fronthaul network, backhaul network, power consumption, function placement

I. INTRODUCTION

Toward the deployment of the 5th generation mobile network [1], various new technologies such as millimeter wave communication [2] and large scale multiple-input and multiple-output [3] are studied and demonstrated. Furthermore, the redesign of Radio Access Network and fronthaul/backhaul network that constitute the mobile network is actively conducted. In particular, Cloud RAN (C-RAN) [4], functional splitting of baseband processing, and the integration of the fronthaul and backhaul networks [5], are being considered based on Software Defined Network [6] and Network Functions Virtualization [7] technologies.

In [8], the problem of the current fronthaul network with deployment of C-RAN have been pointed out, the function splitting between Base Band Unit and Remote Radio Heads (RRHs) have been rethought, and a new fronthaul interface called Next-Generation Fronthaul Interface forwarding packetize data of fronthaul network has been proposed. By applying the virtualization technologies to the fronthaul network, the baseband processing is implemented as a software and deployed on virtualized server pool on the cloud environment. It enables the centralized control of RRHs to use radio resources efficiently and enhance network throughput, as well as reducing server resource utilization.

Also, in [5], integrated control of mobile fronthaul and backhaul network has been proposed. By the integrated control

of the fronthaul and backhaul networks [5] [9], adaptive and dynamic resource allocation to fronthaul and backhaul network functions can be realized. Existing researches showed that by applying such technologies to the mobile network, the communication performance of user equipments, the resource utilization efficiency, and the power consumption can be reduced. However, especially on the integration of the fronthaul and backhaul networks, the quantitative evaluation of such performance improvement has not been conducted in the past literatures.

In this paper, focusing of the function placement problem for both of fronthaul and backhaul networks, the performance of the mobile network is evaluated using mathematical analysis. For this purpose, the analysis model of the mobile network with integrated control of fronthaul and backhaul networks is constructed, considering the effect of traffic volume, packet loss rate, packet transmission latency, and power consumption. Then, numerical results of the analysis model are shown to reveal the advantage of appropriate function placement on the integrated fronthaul/backhaul network in terms of power consumption and network performance.

The remainder of this paper is organized as follows. In Section II, we show the analysis model for evaluating the integrated fronthaul/backhaul networks. In Section III we present numerical evaluation results of the analysis. Finally, in Section IV we present the conclusion of this paper with future work.

II. ANALYSIS MODEL FOR INTEGRATED FRONTHAUL/BACKHAUL NETWORK

A. Network Model

There are N^{NODE} nodes $(n_1, n_2, \dots, n_{N^{NODE}})$ in the network. Let \mathbf{N} be a set of nodes. Node n_i has N_i network interfaces $(f_{i,1}, f_{i,2}, \dots, f_{i,N_i})$. Let \mathbf{F}_{n_i} be the set of network interfaces of node n_i .

A network interface of a node and a network interface of another node connected each other are treated in a network interface pair. We define a link $(f_{i,x}, f_{j,y})$ as a network interface pair, and introduce a network topology \mathbf{T} as a set

of links. Link $(f_{i,x}, f_{j,y})$ has the bandwidth $\beta_{f_{i,x}, f_{j,y}}$ and the propagation delay $\tau_{f_{i,x}, f_{j,y}}$.

For node n_i , we define π_{n_i} as the processing power to process flow packets passing through the node. Note that multiple network functions can be placed on a single node. Network functions considered in this paper are application servers, Evolved Packet Core (EPC) nodes of mobile core networks, BaseBand Unit (BBU), and Optical Network Unit (ONU) and Optical Line Terminal (OLT).

We define a UE as a node that generates a network traffic, while it does not accommodate any other traffic-generating nodes. Also, we define an edge node as a node that directly connected to UEs and accommodating the traffic from the UEs. Furthermore, we denote an upper node as a node other than the UE and the edge node.

B. Traffic Model

In this analysis, it is assumed that traffic requiring a constant bandwidth such as Common Public Radio Interface (CPRI) traffic and best-effort traffic coexist in the network, and that the former is handled with higher priority in the network. Therefore, when a high priority traffic passes through a link, the link bandwidth is utilized for the traffic in a guaranteed manner and the remaining bandwidth is shared by other best-effort traffic. In what follows, the model for best-effort traffic is described.

When traffic from multiple UEs is multiplexed on an output link at the edge node, that directly accommodates the UEs, we consider the detailed traffic characteristics, such as periodical communication and synchronization effect among UEs. On the other hand, regarding traffic multiplexing in upper nodes, we ignore such effects for analysis simplification.

1) *Traffic Demand*: We assume that there are N_{App} applications in the network, and a set of the applications \mathbf{A} is expressed $\mathbf{A} = \{a_1, a_2, \dots, a_{N_{App}}\}$.

We define $D_{s,d,h}$ as a traffic demand of application h ($a_h \in \mathbf{A}$) from node n_s to node n_d ($s \neq d, n_s, n_d \in \mathbf{N}$). When a node has multiple applications that generate network traffic to an identical destination node, we define traffic demands on the same node pair. There is traffic from node s to node d by multiple applications, a traffic demand is set for each application. An application traffic is described as a periodical ON-OFF traffic, which has a constant length of the period. Each period consists of ON and OFF sections. In ON section a source node generate traffic, while it does not generate traffic in OFF section. The length of a period, and that of ON and OFF sections are denoted by T , T^{ON} , and T^{OFF} , respectively ($T = T^{ON} + T^{OFF}$). The traffic demand which generates network traffic at constant bit rate can be described as $T^{ON} = T$ and $T^{OFF} = 0$

The traffic characteristics of application a_h includes T , T^{ON} , T^{OFF} , a traffic rate in the ON section δ_h , an end-to-end delay constraint ω_h , an end-to-end throughput constraint ν_h , and synchronization level σ_h ($0 \leq \sigma_h \leq 1$). The synchronization level means the degree of synchronization of ON sections of the traffic from multiple nodes. When we consider that a

certain communication period starts at 0 and the length of the period of application a_h is T_h , ON sections of the application traffic from multiple nodes starts and ends within the interval of $[0, (1 - \sigma_h)T_h]$. We do not consider the synchronization level of network traffic from different applications.

We define $R_{s,d,h} = \{\mathbf{N}_{s,d,h}^R, \mathbf{I}_{s,d,h}^R\}$ as a route of a network traffic by application a_h from node n_s to node n_d . Note that $\mathbf{N}_{s,d,h}^R$ is defined as a set of nodes passing through the route, and $\mathbf{I}_{s,d,h}^R$ is defined as a set of output network interfaces of the link used in the route.

In what follows, on the traffic by application a_h from network interface $f_{i,j}$ of node n_i , we denote $T_{h,f_{i,j}}$ as the communication period length, $T_{h,(f_{i,j})}^{ON}$ as the length of ON section, $T_{h,(f_{i,j})}^{OFF}$ as the length of OFF section, and $\delta_{h,(f_{i,j})}$ as the traffic rate in the ON section.

2) *Traffic Calculation at Edge nodes*: We first consider the traffic rate calculation at edge nodes which accommodate multiple source nodes with different applications. Since each application has ON and OFF section and the number of possible combinations of traffic generation states of applications are $2^{N_{App}}$. In our model, for analysis simplification, we don't consider the case when the traffic characteristics change in the ON section. Therefore, all application has two state, one is the state of sending traffic, that is, the state in ON section, the other is the state in which no traffic is sent, that is, the state in OFF section. Note that when node n_i does not send the traffic of application a_h at network interface $f_{i,j}$, we regard that the application has on ON section and set $T_{h,f_{i,j}}^{ON} = 0$. We define a set of application state combinations on network interface $f_{i,j}$ as $\mathbf{C}_{f_{i,j}} = \{c_{1,f_{i,j}}, c_{2,f_{i,j}}, \dots, c_{2^{N_{App}}, f_{i,j}}\}$, where $c_{k,f_{i,j}}$ consists of the state of all application states, denoted by $s_{h,f_{i,j}}^k$. Then, c_k and $s_{h,f_{i,j}}^k$ can be expressed as follows.

$$c_{k,f_{i,j}} = (s_{1,f_{i,j}}^k, s_{2,f_{i,j}}^k, \dots, s_{2^{N_{App}}, f_{i,j}}^k) \quad (1)$$

$$s_{h,f_{i,j}}^k = \begin{cases} 1 & \text{if } \exists D_{s,d,h}, \mathbf{I}_{s,d,h}^R \ni f_{i,j} \text{ and } a_h \text{ is ON} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

When a node is a UE that has a single application, or when a node accommodates only one such UE, the traffic characteristics on the network interface of the node is identical of the traffic characteristics of the application. Therefore, we can calculate $\delta_{h,f_{i,j}}$, $T_{h,f_{i,j}}$, and $T_{h,f_{i,j}}^{ON}$ of such node as follows.

$$\delta_{h,f_{i,j}} = \begin{cases} \delta_h & \text{if } \exists D_{s,d,h}, \mathbf{I}_{s,d,h}^R \ni f_{i,j} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

$$T_{h,f_{i,j}} = T_h \quad (4)$$

$$T_{h,f_{i,j}}^{ON} = T_h^{ON} \quad (5)$$

$$T_{h,f_{i,j}}^{OFF} = T_h - T_h^{ON} = T_h^{OFF} \quad (6)$$

On the other hand, when a node accommodate multiple UE nodes which have one and the same application, the traffic characteristics on the outgoing network interface can be calculated as follows.

$$T_{h,f_{i,j}} = T_h \quad (7)$$

$$T_{h,f_{i,j}}^{OFF} = \sigma_h T_{h,f_{i,j}} \quad (8)$$

$$T_{h,f_{i,j}}^{ON} = (1 - \sigma_h) T_{h,f_{i,j}} \quad (9)$$

$$\delta_{h,f_{i,j}} = \begin{cases} \sum_{\mathbf{I}_{s,d,h}^R \ni f_{i,j}} \left(\delta_h \frac{T_h^{ON}}{T_{h,f_{i,j}}^{ON}} \right) & \text{if } \exists D_{s,d,h}, \mathbf{I}_{s,d,h}^R \ni f_{i,j} \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

We next consider the case where a node accommodate multiple UE nodes with multiple applications. At the network interface $f_{i,j}$ of node n_i , we define $r_{f_{i,j}}^k$ as the ratio of the time when the combination of the application state is $c_{k,f_{i,j}}$, and $r_{h,f_{i,j}}^k$ as the ratio of the time when the state of application a_h is $s_{h,f_{i,j}}^k$, $r_{h,f_{i,j}}^k$ and $r_{f_{i,j}}^k$ are calculated as follows.

$$r_{h,f_{i,j}}^k = \begin{cases} \frac{T_{h,f_{i,j}}^{ON}}{T_{h,f_{i,j}}} & \text{if } s_{h,f_{i,j}}^k = 1 \\ 1 & \text{if } \exists D_{s,d,h}, \mathbf{I}_{s,d,h}^R \ni f_{i,j} \\ 1 - \frac{T_{h,f_{i,j}}^{ON}}{T_{h,f_{i,j}}} & \text{otherwise} \end{cases} \quad (11)$$

$$r_{f_{i,j}}^k = \prod_{h \in \mathbf{A}} r_{h,f_{i,j}}^k \quad (12)$$

At a certain state combination of applications, the average traffic rate from multiple applications with different traffic characteristics is calculated as a sum of average traffic rate from each application which generate the traffic in the state combination. Then, $\delta_{f_{i,j}}^k$ is calculated as follows.

$$\delta_{f_{i,j}}^k = \sum_h \delta_{h,f_{i,j}} \cdot s_{h,f_{i,j}}^k \quad (13)$$

3) *Traffic Calculation at Upper Nodes:* For analysis simplification, we do not consider the detailed traffic generation states from applications at upper nodes where traffic from multiple edge nodes are multiplexed. Then, the traffic rate at network interface $f_{i,j}$ on such node can be calculated as follows.

$$\delta_{f_{i,j}} = \sum_{\mathbf{I}_{s,d,h}^R \ni f_{i,j}} \left(\delta_h \frac{T_h^{ON}}{T_h} \right) \quad (14)$$

C. Power Consumption Model

1) *Power Consumption at Node:* The power consumption of a node is determined by the sum of the power consumption of network interfaces of the node and the power consumption for node processing, meaning executing network functions to the traffic at the node. Therefore, the power consumption of node e_{n_i} can be calculated as follows, where $e_{f_{i,j}}$ is the power consumption of the network interface $f_{i,j}$ of node n_i and $e_{n_i}^{Proc}$ is the power consumption for node processing.

$$e_{n_i} = e_{n_i}^{Proc} + \sum_j e_{f_{i,j}} \quad (15)$$

2) *Energy Proportionality:* We utilize the energy proportionality model presented in the article [10] for determining the power consumption of network interfaces and node processing. The graph in Figure 1 explains the energy proportional model utilized in this paper, where x axes is the rate of the traffic at which the network interface/node process and y axes is the power consumption. In this graph we show the case for ideal power consumption in green line, actual power consumption in blue line, and $E(x)$ used in this paper in orange line.

In ideal power consumption, the power consumption becomes zero when there is no traffic processed. However, in the actual case, even when there is no network traffic on the node, some amount of power is required for activate the node device. Furthermore, especially for network nodes, additional hardwares such as interface boards should be installed for processing larger amount of traffic rate, requiring additional constant power consumption.

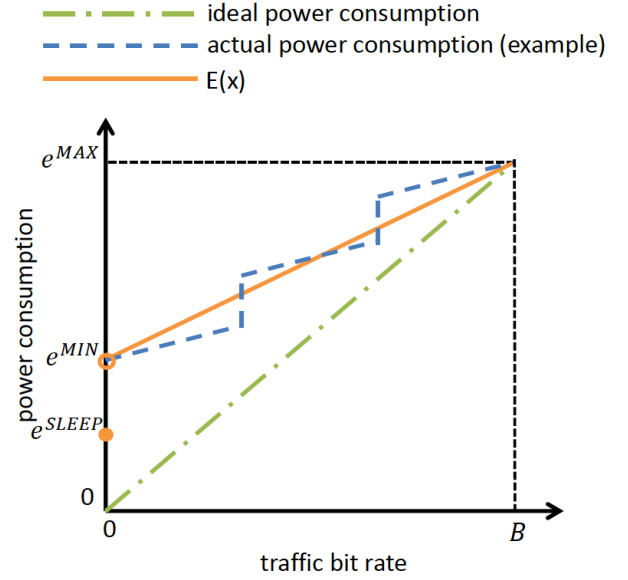


Fig. 1. Energy proportionality of devices

3) *Power Consumption of Network Interfaces:* We utilize the following equations for determining the power consumption of network interface $f_{i,j}$:

$$E_{f_{i,j}}(x) = \begin{cases} x \frac{e_{f_{i,j}}^{MAX} - e_{f_{i,j}}^{MIN}}{B_{f_{i,j}}} + e_{f_{i,j}}^{MIN} & x > 0 \\ e_{f_{i,j}}^{SLEEP} & x = 0 \end{cases} \quad (16)$$

where x is the traffic rate to be processed, $e_{f_{i,j}}^{MAX}$ is the power consumption when the network interface process the network traffic at the maximum capacity, and $e_{f_{i,j}}^{MIN}$ is the power consumption when there is no network traffic. $e_{f_{i,j}}^{SLEEP}$ is the power consumption when the network interface is at the sleep mode when supported.

When node n_i is an edge node, $e_{f_{i,j}}$ is calculated as follows, by using the state combinations.

$$e_{f_{i,j}} = \sum_{c_{k,f_{i,j}}} r_{f_{i,j}}^k E_{f_{i,j}}(\delta_{f_{i,j}}^k) \quad (17)$$

On the other hand, when node n_i is an upper node, $e_{f_{i,j}}$ is calculated as follows.

$$e_{f_{i,j}} = E_{f_{i,j}}(\delta_{f_{i,j}}) \quad (18)$$

4) *Power Consumption for Node Processing:* As in the network interfaces, the power consumption for node processing is calculated based on the energy proportionality model, as shown below:

$$E_{n_i}(x) = \begin{cases} x \frac{e_{n_i}^{MAX} - e_{n_i}^{MIN}}{B_{n_i}} + e_{n_i}^{MIN} & x > 0 \\ e_{n_i}^{SLEEP} & x = 0 \end{cases} \quad (19)$$

where x is the traffic rate to be processed, $e_{n_i}^{MAX}$ is the power consumption when the node process the network traffic at the maximum capacity, and $e_{n_i}^{MIN}$ is the power consumption when there is no network traffic to be processed. $e_{n_i}^{SLEEP}$ is the power consumption when the node is at the sleep mode when

supported. The traffic rate x is obtained by considering the characteristics of edge nodes and upper nodes.

$$x = \begin{cases} \sum_j \sum_{c_{k,f_{i,j}}} \delta_{f_{i,j}}^k r_{f_{i,j}}^k & \text{if Node } n_i \text{ is EdgeNode or UE} \\ \sum_j \delta_{f_{i,j}} & \text{otherwise} \end{cases} \quad (20)$$

D. Latency and Packet Loss Rate

The end-to-end latency of the application traffic is the sum of propagation delay of links on the path between the source node and the destination node, the queuing delay at the network interfaces, and the processing delay at intermediate nodes. The second and third ones are obtained based on the simple queuing theory.

We assume that packet losses may occur at network interfaces on the path. The packet loss rate is also obtained through queuing model.

1) *Queueing Delay and Packet Loss Rate at Network Interfaces:* We exploit M/M/1/K queuing model [11] for determining the queuing delay and packet loss rate at network interfaces. For edge nodes, the queuing delay $w_{f_{i,j}}^k$ and the packet loss rate $p_{f_{i,j}}^k$ of the network interface $f_{i,j}$ on the state combination $c_{k,f_{i,j}}$ are obtained by using the following equations.

$$\lambda_{f_{i,j}}^k = \frac{\delta_{f_{i,j}}^k}{m} \quad (21)$$

$$\mu_{f_{i,j}}^k = \frac{\beta_{f_{i,j}}}{m} \quad (22)$$

$$\rho_{f_{i,j}}^k = \frac{\lambda_{f_{i,j}}^k}{\mu_{f_{i,j}}^k} \quad (23)$$

$$p_{f_{i,j}}^k = \frac{\rho_{f_{i,j}}^k K_{f_{i,j}}}{1 + \rho_{f_{i,j}}^k + \dots + \rho_{f_{i,j}}^k K_{f_{i,j}}} \quad (24)$$

$$p_{0_{f_{i,j}}}^k = \begin{cases} \frac{1 - \rho_{f_{i,j}}^k}{1 - \rho_{f_{i,j}}^k K_{f_{i,j}+1}} & (\rho_{f_{i,j}}^k \neq 1) \\ \frac{1}{K_{f_{i,j}} + 1} & (\rho_{f_{i,j}}^k = 1) \end{cases} \quad (25)$$

$$L_{q_{f_{i,j}}}^k = \begin{cases} \frac{\rho_{f_{i,j}}^k}{1 - \rho_{f_{i,j}}^k} - \frac{\rho_{f_{i,j}}^k (K_{f_{i,j}} \rho_{f_{i,j}}^k + 1)}{1 - \rho_{f_{i,j}}^k K_{f_{i,j}+1}} & (\rho_{f_{i,j}}^k \neq 1) \\ \frac{K(K-1)}{2(K+1)} & (\rho_{f_{i,j}}^k = 1) \end{cases} \quad (26)$$

$$L_{f_{i,j}}^k = L_{q_{f_{i,j}}}^k + 1 - p_{0_{f_{i,j}}}^k \quad (27)$$

$$w_{f_{i,j}}^k = \frac{L_{f_{i,j}}^k}{\lambda_{f_{i,j}}^k (1 - p_{f_{i,j}}^k)} \quad (28)$$

We utilized the similar equations for upper nodes. Specifically, it is expressed by an expression expect the suffix k in the above equations 21–28.

2) *Processing Time at Node:* For determining the node processing delay, the M/G/1/PS queuing model with r of parallelism parameter is exploited. With the job arriving rate of λ , workload distribution of $S(x)$ and its mean value of $E[S]$, and system utilization of $\rho = \lambda E(S)$, the mean response time, $E[R]$ is obtained as follows.

$$E[R] = \frac{\rho^r}{1 - \rho} \frac{E[S^2]}{2E[S]} + \frac{1 - \rho^r}{1 - \rho} E[S] \quad (29)$$

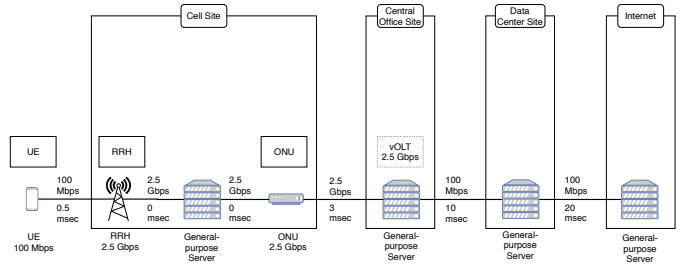


Fig. 2. Network environment

For determining the job arrival rate at the node n_i , denoted by λ_{n_i} , we utilized the following calculations:

$$\lambda_{n_i} = \frac{x}{m} \quad (30)$$

$$x = \begin{cases} \sum_j \sum_{c_{k,f_{i,j}}} \delta_{f_{i,j}}^k r_{f_{i,j}}^k & \text{if Node } n_i \text{ is EdgeNode or UE} \\ \sum_j \delta_{f_{i,j}} & \text{otherwise} \end{cases} \quad (31)$$

where m is the mean packet size of the application traffic.

III. NUMERICAL EVALUATION RESULTS AND DISCUSSIONS

A. Evaluation Environment

1) *Network Environment:* Figure 2 depicts the network environment for numerical evaluations in this section. The network consists of four network sites, which are a cell site, a central office site, a data center site, and the Internet site. The four sites are interconnected in a line topology. The link bandwidth between the cell site and the central office site, that between the central office site and the data center site, and that between the data center site and the Internet site are set to 2.5 Gbps, 100 Mbps, and 100 Mbps, respectively. The propagation delays are configured to 3 msec, 10 msec, and 20 msec. For simplicity, there is only one UE connected to RRH at the cell site via a wireless network. The wireless network has 100 Mbps capacity and 0.5 msec propagation delay.

Each site has physical servers for hosting virtual machines to deploy the network functions such as vBBU, vEPCs, and application servers. The servers at the cell site and the central office site have the capacity of 2.4 Gbps, assuming to process CPRI traffic between RRH and vBBU. The servers at the data center site and the Internet site has 100 Mbps capacity for processing vEPCs and application servers.

2) *Power Consumption Setting:* The physical servers at the four sites consume 200 W, and additional 20 W is required for hosting a virtual server that deploys a network function. Unlike the network interface, we assume that the physical and virtual servers consume the power constantly regardless of the processing traffic rate.

3) *Application Traffic Setting:* It is assumed that the UE generate network traffic of two applications (application 1 and application 2) to application servers. The network traffic from application 1 we assume the traffic for updating the current location of a car at the regular intervals for autonomous

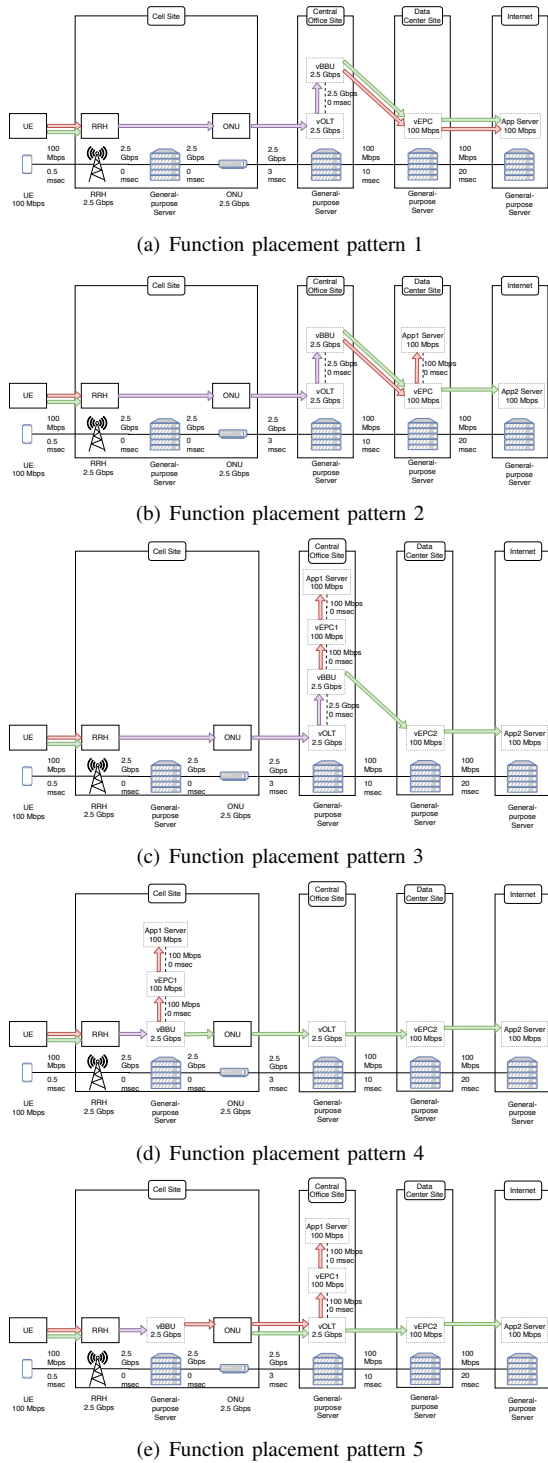


Fig. 3. Function placement patterns for performance evaluation

driving. Therefore, the traffic is generated at the regular intervals and it has a strict latency constraint. We set the communication cycle of application 2 is set to 1 sec, and the data amount to be transmitted in each cycle is set to 125 KBytes. In the evaluation we change the duration of the data transmission in each cycle, that is T^{ON} in the analysis model

to investigate the effect of the bursty nature on the application performance.

Traffic from application 2 is generated at fixed bit rate that does not have explicit latency constraint, assuming the video traffic from drive recorder's camera on a car. In the evaluation, the data rate is changed from 1 Mbps to 100 Mbps and assess the effect of the traffic rate on the network and application performance.

We also consider the CPRI traffic between RRH and vBBU for C-RAN configuration in Figure 2. In the evaluation by using the analysis model in Section II, we utilize a traffic demand between the UE and RRH, and an traffic demand between vBBU and the application server, as well as the traffic demand of CPRI between RRH and vBBU.

When a network function process the network traffic of both applications at a single site, we deploy one virtual machine of the network function for the both applications. On the other hand, when a network function process the network traffic of both applications at different sites, we deploy one virtual machine of the network function for each application. For example, when the network function for an application with tight delay constraint is located at the site near the UE, we need an additional virtual machine as well as increased power consumption.

4) *Function Placement Patterns*: For accommodating the traffic demands in Subsection III-A3 on the network in Figure 2, we consider the various placement patterns of network functions (vBBU, vEPC, and application servers) for application 1 and application 2. Figure 3(a) – 3(e) depicts the five patterns of the network function placement. In these figures, the red, green, and purple arrows represents the traffic flows of application 1, 2, and CPRI, respectively.

Pattern1:

The traffic from both applications are processed a common vEPC at the data center site, and a common application server at the Internet site. vBBU for baseband processing are deployed at the central office. (Figure 3(a))

Pattern2:

The application server for application 1 is located at the data center site. (Figure 3(b))

Pattern3:

The application server and vEPC for application 1 are located at the central office site. (Figure 3(c))

Pattern4:

vBBU for baseband processing are deployed at the cell site. The application server and vEPC are also located at the cell site. (Figure 3(d))

Pattern5:

vBBU for baseband processing are deployed at the cell site. The application server and vEPC are located at the central office site. (Figure 3(e))

B. Results and Discussions

We assess the effect of function placement for both applications explained in Subsection III-A3. As explained in

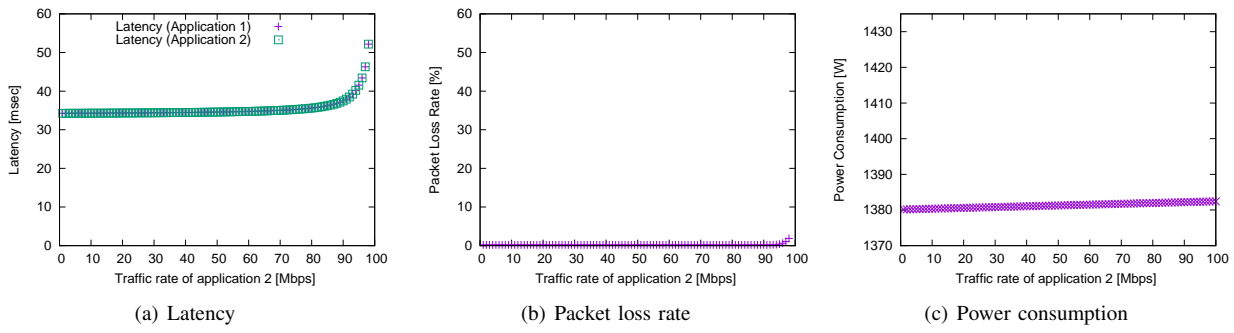


Fig. 4. Performance evaluation results for placement pattern 1

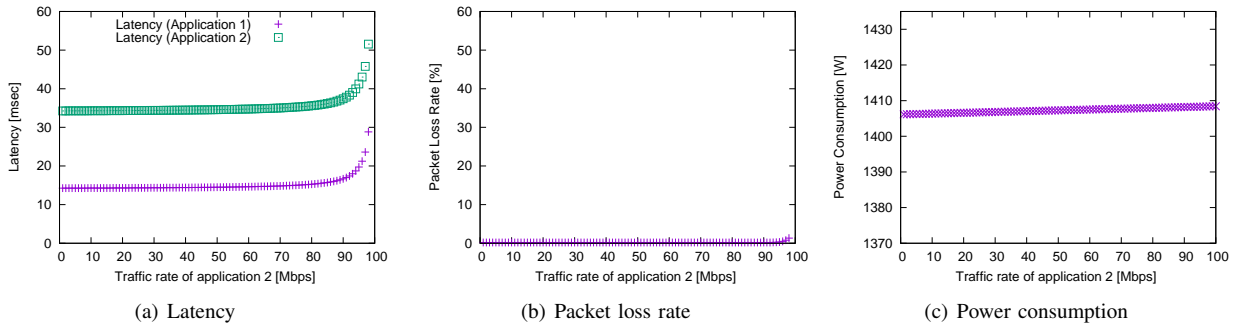


Fig. 5. Performance evaluation results for placement pattern 2

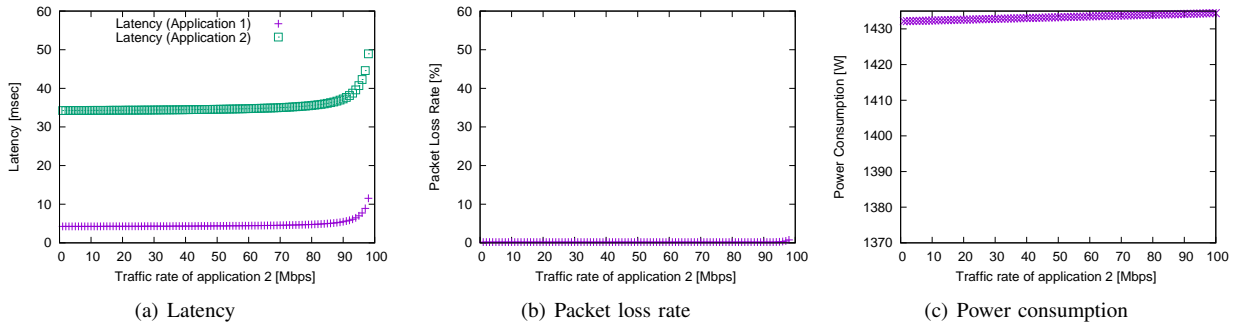


Fig. 6. Performance evaluation results for placement pattern 3

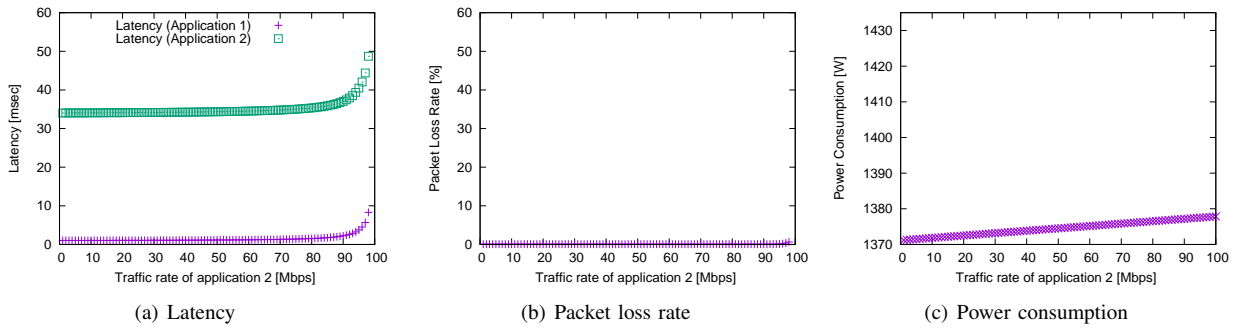


Fig. 7. Performance evaluation results for placement pattern 4

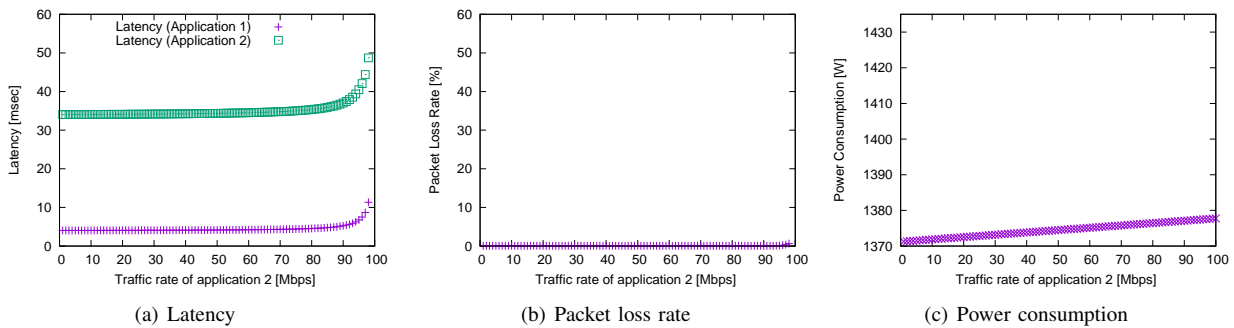


Fig. 8. Performance evaluation results for placement pattern 5

Subsection III-A4, we compare five patterns of function placement depicted in Figures 3(a) – 3(e). Figures 4 – 8 show the evaluation results of patterns 1– 5, respectively. We set the communication period of application 1 to 1 sec and the length of ON section is identical to the communication period, meaning that the application traffic is transmitted at constant bit rate of 1 Mbps.

By comparing pattern 1 in Figure 4 and pattern 2 in Figure 5, we can observe that the latency of application 2 decreases by moving the application server for application 2 is located from the Internet site to the data center site. On the other hand, the power consumption increases in pattern 2. This is because additional power consumption by a virtual server for application 1 is required at the data center site. We confirmed that the power consumption of the network interfaces slightly decreases by removing the network traffic of application 2 between the data center site and the Internet site. However, the increase by additional virtual server is quite larger than the decrease by shortening the network path.

From Figure 6, we can also confirm the same effect by moving servers for application 1 to further nearer location to UEs, where the application server and vEPC for application 1 are located at the central office site. As in the above case, we have decreased latency for application 1 at the sacrifice of increased total power consumption.

In pattern 4, whose evaluation results are presented in Figure 7, the latency of application 1 is further decreased since the application server and vEPC are located at the cell site, which is the nearest site from UEs. However, the degree of the latency decrease is not so large because the propagation delay between the cell site and the central office site is small. The advantage of pattern 4 can be observed in power consumption found in Figure 7(c). This is because the amount of traffic between the cell site and the central office is greatly decreased by moving BBU function from the central office to the cell site.

Comparing with pattern 4, pattern 5 in Figure 8 has almost no advantage since the latency of application 1 becomes slightly large by placing the application server and vEPC are located at the central office site. On the other hand the advantage of the small power consumption can be achieved by placing BBU function at the cell site. In the actual situation, however, the power consumption efficiency and the server pool capacity of physical and virtual machines may be different at the cell site and the central office site, that affects the total power consumption especially when the numbers of UEs and RRHs increase. The evaluation of such cases is one of important future work.

IV. CONCLUSION

In this paper, we constructed the mathematical analysis model for evaluating the performance of the integrated fronthaul/backhaul network. The analysis method is based on the simple but effective queueing theory and it can treat the effects as follows: the traffic characteristics of applications such as periodical transmission; flexible network function placement

strategies realized by SDN and NFV technologies; and power consumption characteristics of network interfaces and packet processing on network nodes and servers.

The effectiveness of the proposed analysis model was confirmed by numerical evaluation results assuming simple network environment with single UE, and presented that the function placement greatly affect the power consumption of the whole network system and the application performance.

For future work, we plan to evaluate the integrated fronthaul/backhaul network in larger-scale in terms of the number of applications, UEs, and network nodes. In addition, to obtain more reliable numerical results, we need to refine parameter settings of the power consumption characteristics of network interfaces and physical/virtual servers.

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, pp. 1065–1082, June 2014.
- [2] Z. Pi and F. Khan, "An Introduction to Millimeter-Wave Mobile Broadband Systems," *IEEE Communications Magazine*, vol. 49, pp. 101–107, June 2011.
- [3] D. Gesbert, M. Shafi, D. shan Shiu, P. J. Smith, and A. Naguib, "From Theory to Practice: An Overview of MIMO Space-Time Coded Wireless Systems," *IEEE Journal on Selected Areas in Communications*, vol. 21, pp. 281–302, Apr. 2003.
- [4] 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, pp. 405–426, Firstquarter 2015.
- [5] A. de la Oliva, X. C. Pérez, A. Azcorra, A. D. Giglio, F. Cavaliere, D. Tiegelbekkers, J. Lessmannm, T. Haustein, A. Mourad, and P. Iovanna, "Xhaul: Toward an Integrated Fronthaul/Backhaul Architecture in 5G Networks," *IEEE Wireless Communications*, Oct. 2015.
- [6] D. Kreutz, F. M. V. Ramos, P. E. Verissimo, C. E. Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-Defined Networking: A Comprehensive Survey," *Proceedings of the IEEE*, vol. 103, pp. 14–76, Jan. 2015.
- [7] R. Mijumbi, J. Serrat, J. L. Gorricho, N. Bouten, F. D. Turck, and R. Boutaba, "Network Function Virtualization: State-of-the-Art and Research Challenges," *IEEE Communications Surveys Tutorials*, vol. 18, pp. 236–262, Firstquarter 2016.
- [8] C.-L. I, Y. Yuan, J. Huang, S. Ma, C. Cui, and R. Duan, "Rethink Fronthaul for Soft RAN," *Communications Magazine, IEEE*, 2015.
- [9] X. Costa-Perez, A. Garcia-Saavedra, X. Li, T. Deiss, O. Delgado, A. Di Giglio, A. Mourad, *et al.*, "5G-Crosshaul: an SDN/NFV Integrated Fronthaul/Backhaul Transport Network Architecture," *IEEE wireless communications*, Feb. 2017.
- [10] P. Mahadevan, P. Sharma, S. Banerjee, and P. Ranganathan, *A Power Benchmarking Framework for Network Devices*, pp. 795–808. Springer Berlin Heidelberg, May 2009.
- [11] D. Gross, J. F. Shortle, J. M. Thompson, and C. M. Harris, *Fundamentals of Queueing Theory, 4th Edition*. John Wiley and Sons, Inc., 2008.