Master's Thesis

Title

Power Consumption Analysis of Cloud-based Integrated Mobile Fronthaul/Backhaul Network

Supervisor

Professor Morito Matsuoka

Author

Rina Yamasaki

February 6th, 2019

Graduate School of Information Science and Technology

Osaka University

Master's Thesis

Power Consumption Analysis of
Cloud-based Integrated Mobile Fronthaul/Backhaul Network

Rina Yamasaki

Abstract

Toward the deployment of the 5th generation mobile network, various new technologies such as millimeter wave communication and large scale multiple-input and multiple-output are studied and demonstrated. Furthermore, the redesign of radio access networks, fronthaul network, and backhaul network that constitute the mobile network is actively conducted. In particular, Cloud RAN, functional splitting of baseband processing, network slicing, and the integration of the fronthaul and backhaul networks, are being considered based on software defined networking and network functions virtualization technologies.

By applying the virtualization technologies to the fronthaul network, the baseband processing is implemented as a software and deployed on virtualized server pool in the cloud environment. It enables the centralized control of remote radio heads to urilize radio resources efficiently and enhance network throughput, as well as reducing server resource urilization. Also, by the integrated control of the fronthaul and backhaul networks, 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 thesis, focusing of the integrated control of the 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, server/link resource limitation, and their power consumption characteristics. Then, numerical results of the analysis model are shown to

reveal the advantage of the integrated fronthaul/backhaul network in terms of power consumption and network performance. Specifically, the importance of the placement of fronthaul and backhaul network functions is presented.

Numerical evaluation results show that the burstiness of the network traffic can affect the application performance largely, and that by carefully placing the network functions for delay-sensitive application traffic, the end-to-end latency can decreased by about 97 % at little sacrifice of the power consumption.

Keywords

5th Generation Mobile Network
Integrated Fronthaul/Backhaul Network
Software Defined Networking (SDN)
Network Functions Virtualization (NFV)
Cloud-Radio Access Network (C-RAN)
Power Consumption Analysis

Contents

List of Figures

List of Tables

1 Introduction

Toward the deployment of the 5th generation mobile network [1, 2], various new technologies such as millimeter wave communication [3] and large scale multiple-input and multiple-output [4] 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) [5–7], functional splitting of baseband processing, network slicing [8, 9] and the integration of the fronthaul and backhaul networks [10, 11], are being considered based on Software Defined Network [12, 13] and Network Functions Virtualization [14, 15] technologies.

In [5], 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 urilization.

Also, in [10], integrated control of mobile fronthaul and backhaul network has been proposed. By the integrated control of the fronthaul and backhaul networks [10, 11], 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 thesis, focusing of the integrated control of the 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 the integrated fronthaul/backhaul network in terms of power consumption and

network performance. Specifically, the importance of the placement of fronthaul and backhaul network functions is presented.

In Section 2, the analysis model which consist of the network model, the traffic model, the power consumption model, and the packet transmission latency model are presented. Section 3, numerical results of the analysis model are shown to reveal the advantage of the integrated fronthaul/backhaul network in terms of power consumption and network performance. Finally, in Section 4, We conclude this thesis with a brief summary and an outline of future work.

2 Analysis Model for Integrated Fronthaul/Backhaul Network

2.1 Integrated Fronthaul/Backhaul Network

Figure 1 depicts the example of the mobile network where fronthaul network and backhaul network are integrated based on SDN and NFV technology. In Figure 1, the network consists of User Equipements (UE), application servers, Mobility Management Entities (MME), Home Subscriber Server (HSS), Serving Gateways (SGWs), Packet data network Gateways (PGWs), evolved NodeBs (eNodeBs), Optical Network Units (ONUs) and Optical Line Terminals (OLTs) for constructing Passive Optical Networks (PON), Baseband Units (BBUs), and Remote Radio Heads (RRHs).

MME, HSS, SGW, PGW, and eNodeB are components of Evolved Packet Core (EPC) nodes of mobile core networks. In this thesis, we consider that the mobile core network is a part of backhaul. In Figure 1, they are presented in green. ONU, OLT, BBU, and RRH are components of mobile fronthaul network that is presented in orange.

In the traditional mobile network, the function placement for fronthaul and backhaul networks are tightly dependent on physical network facility, and these two networks are clearly separated. However, in the integrated fronthaul/backhaul networks, the network function are decoupled from physical facility and these functions share the under-layer physical networks. So, flexible and elastic function placement can be realized.

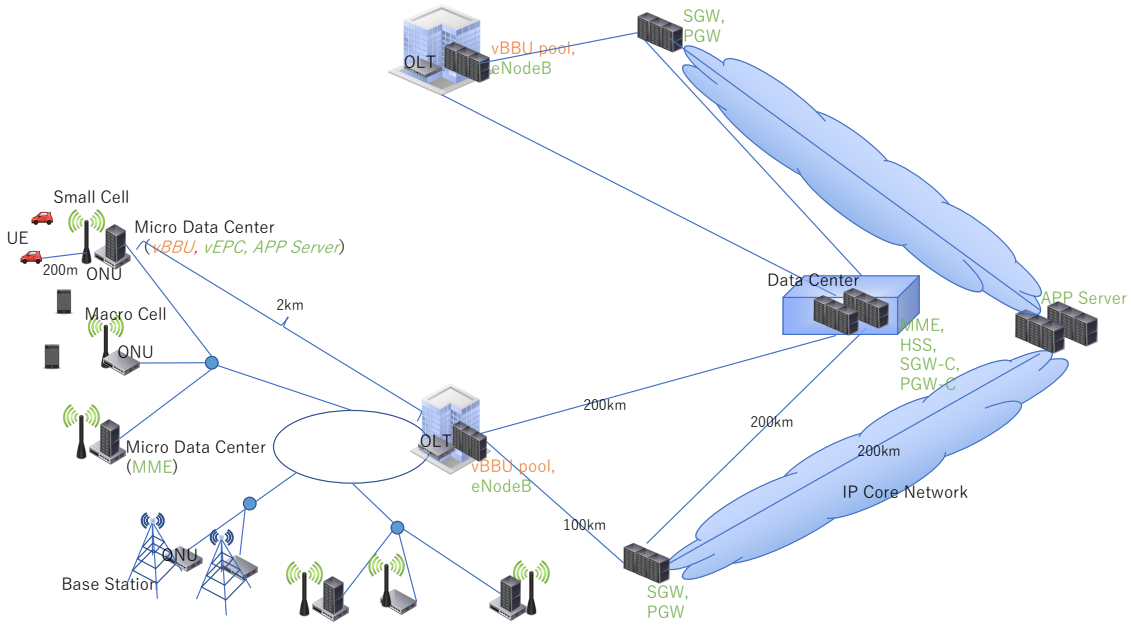


Figure 1: Example of integrated fronthaul and backhaul network

2.2 Network Model

There are N^{NODE} nodes $(n_1, n_2, ..., n_{N^{NODE}})$ in the network. Let $\mathbf{N}$ be a set of nodes. Node n_i has N_i network interfaces $(f_{i,1}, f_{i,2}, ..., f_{i,N_i})$. 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. A network interface on a node may be paired with multiple network interfaces on multiple nodes, such as wireless LAN interfaces between an access point and multiple stations. For example, when node n_1 is an access point of a wireless LAN and accommodates nodes n_2 and n_3, the pairs of network interfaces are $(f_{1,1}, f_{2,1})$ and $(f_{1,1}, f_{3,1})$. 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 thesis are application servers, EPC nodes of mobile core networks, BBU, and ONU and 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.

2.3 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.

2.3.1 Traffic Demand

We assume that there are N_{App} applications in the network, and a set of the applications $\mathbf{A}$ is expressed $\mathbf{A} = \{a_1, a_2, ..., a_{N_App}\}$.

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 trrafic 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 ι_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 stats 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 $R_{s,d,h} = \{\mathbf{N}^R_{s,d,h}, \mathbf{I}^R_{s,d,h}\}$ as a route of a network traffic by application a_h from node n_s to node n_d. Note that $\mathbf{N}^R_{s,d,h}$ is defined as a set of nodes passing through the route, and $\mathbf{I}^R_{s,d,h}$ 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, $T^{ON}_{h,(f_{i,j})}$ as the length of ON section, $T^{OFF}_{h,(f_{i,j})}$ as the length of OFF section, and $\delta_{h,(f_{i,j})}$ as the traffic rate in the ON section.

2.3.2 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 $T^{ON}_{h,f_{i,j}} = 0$. We define a set of application state combinations on network interface $f_{i,j}$ as $\mathbf{C}_{f_{i,j}} = \{c_{1,f_{i,j}}, c_{2,f_{i,j}}, \ldots, c_{2^{N_{App}},f_{i,j}}\}$, where $c_{k,f_{i,j}}$ consists of the state of all application states, denoted by $s^k_{h,f_{i,j}}$.Then, c_k and $s^k_{h,f_{i,j}}$ can be expressed as follows.

$$c_{k,f_{i,j}} = (s^k_{1,f_{i,j}}, s^k_{2,f_{i,j}}, \ldots, s^k_{N_{App},f_{i,j}}) \tag{1}$$

$$s^k_{h,f_{i,j}} = \begin{cases} 1 & \text{if } \exists D_{s,d,h}, \mathbf{I}^R_{s,d,h} \ni f_{i,j} \text{ and } a_h \text{ is ON} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

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 $T^{ON}_{h,f_{i,j}}$ of such node as follows.

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

$$T_{h,f_{i,j}} = T_h \tag{4}$$

$$T^{ON}_{h,f_{i,j}} = T^{ON}_h \tag{5}$$

$$T^{OFF}_{h,f_{i,j}} = T_h - T^{ON}_h = T^{OFF}_h \tag{6}$$

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 fol-

lows.

$$T_{h,f_{i,j}} = T_h \tag{7}$$

$$T^{OFF}_{h,f_{i,j}} = \sigma_h T_{h,f_{i,j}} \tag{8}$$

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

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

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 $r^k_{f_{i,j}}$ as the ratio of the time when the combination of the application state is $c_{k,f_{i,j}}$, and $r^k_{h,f_{i,j}}$ as the ratio of the time when the state of application a_h is $s^k_{h,f_{i,j}}$, $r^k_{h,f_{i,j}}$ and $r^k_{f_{i,j}}$ are calculated as follows.

$$r^k_{h,f_{i,j}} = \begin{cases} \frac{T^{ON}_{h,f_{i,j}}}{T_{h,f_{i,j}}} & \text{if } s^k_{h,f_{i,j}} = 1 \\ 1 & \text{if } \exists D_{s,d,h}, \mathbf{I}^R_{s,d,h} \not\ni f_{i,j} \\ 1 - \frac{T^{ON}_{h,f_{i,j}}}{T_{h,f_{i,j}}} & \text{otherwise} \end{cases} \tag{11}$$

$$r^k_{f_{i,j}} = \Pi_{h \in \mathbf{A}} r^k_{h,f_{i,j}} \tag{12}$$

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, $\delta^k_{f_{i,j}}$ is calculated as follows.

$$\delta^k_{f_{i,j}} = \sum_h \delta_{h,f_{i,j}} \cdot s^k_{h,f_{i,j}} \tag{13}$$

2.3.3 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.

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

2.4 Power Consumption Model

2.4.1 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.

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

2.4.2 Energy Proportionality

We utilze the energy proportionality model presented in the article [16] for determining the power consumption of network interfaces and node processing. The graph in Figure 2 explains the energy proportional model utilized in this thesis, where x axies is the rate of the traffic at which the network interface/node process and y axies 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 thesis 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.

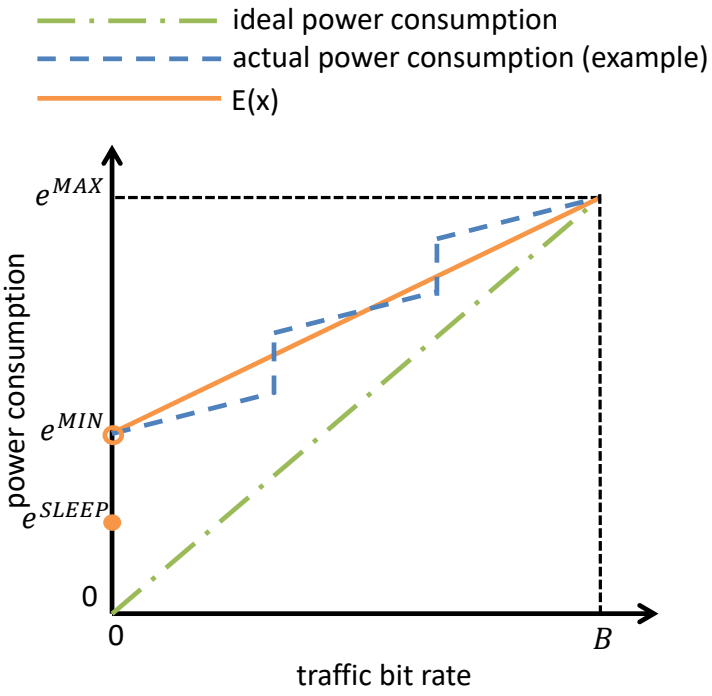


Figure 2: Energy proportionality of devices

2.4.3 Power Consumption of Network Interfaces

We utilize the following equations for determining the power consumption of network interface $f_{i,j}$:

$$\mathrm{E}_{f_{i,j}}(x) = \begin{cases} x\dfrac{e^{MAX}_{f_{i,j}} - e^{MIN}_{f_{i,j}}}{B_{f_{i,j}}} + e^{MIN}_{f_{i,j}} & x > 0 \\ e^{SLEEP}_{f_{i,j}} & x = 0 \end{cases} \tag{16}$$

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

$$e_{f_{i,j}} = \sum_{c_{k,f_{i,j}}} r^k_{f_{i,j}} \mathrm{E}_{f_{i,j}}\left(\delta^k_{f_{i,j}}\right) \tag{17}$$

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

$$e_{f_{i,j}} = \mathrm{E}_{f_{i,j}}(\delta_{f_{i,j}}) \tag{18}$$

2.4.4 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:

$$\mathrm{E}_{n_i}(x) = \begin{cases} x\dfrac{e^{MAX}_{n_i} - e^{MIN}_{n_i}}{B_{n_i}} + e^{MIN}_{n_i} & x > 0 \\ e^{SLEEP}_{n_i} & x = 0 \end{cases} \tag{19}$$

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

$$x = \begin{cases} \sum_{j} \sum_{c_{k,f_{i,j}}} \delta^{k}_{f_{i,j}} r^{k}_{f_{i,j}} & \text{if Node } n_i \text{ is EdgeNode or UE} \\ \sum_{j} \delta_{f_{i,j}} & \text{otherwise} \end{cases} \tag{20}$$

2.5 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 queueing delay at the network interfaces, and the processing delay at intermediate nodes. The second and third ones are obtained based on the simple queueing theory.

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

2.5.1 Queueing Delay and Packet Loss Rate at Network Interfaces

We exploit M/M/1/K queueing model [17] for determining the queueing delay and packet loss rate at network interfaces. For edge nodes, the queueing delay $w^k_{f_{i,j}}$ and the packet loss rate $p^k_{f_{i,j}}$ of the network interface $f_{i,j}$ on the state combination $c_{k,f_{i,j}}$ are obtained by using the following equations.

$$\lambda^k_{f_{i,j}} = \frac{\delta^k_{f_{i,j}}}{m} \tag{21}$$

$$\mu^k_{f_{i,j}} = \frac{\beta_{f_{i,j}}}{m} \tag{22}$$

$$\rho^k_{f_{i,j}} = \frac{\lambda^k_{f_{i,j}}}{\mu^k_{f_{i,j}}} \tag{23}$$

$$p^k_{f_{i,j}} = \frac{{\rho^k_{f_{i,j}}}^{K_{f_{i,j}}}}{1 + \rho^k_{f_{i,j}} + \cdots + {\rho^k_{f_{i,j}}}^{K_{f_{i,j}}}} \tag{24}$$

$${p_0}^k_{f_{i,j}} = \begin{cases} \dfrac{1 - \rho^k_{f_{i,j}}}{1 - {\rho^k_{f_{i,j}}}^{K+1}} & (\rho^k_{f_{i,j}} \neq 1) \\ \dfrac{1}{K+1} & (\rho^k_{f_{i,j}} = 1) \end{cases} \tag{25}$$

$${L_q}^k_{f_{i,j}} = \begin{cases} \dfrac{\rho^k_{f_{i,j}}}{1 - \rho^k_{f_{i,j}}} - \dfrac{\rho^k_{f_{i,j}}(K{\rho^k_{f_{i,j}}}^K + 1)}{1 - {\rho^k_{f_{i,j}}}^{K+1}} & (\rho^k_{f_{i,j}} \neq 1) \\ \dfrac{K(K-1)}{2(K+1)} & (\rho^k_{f_{i,j}} = 1) \end{cases} \tag{26}$$

$$L^k_{f_{i,j}} = {L_q}^k_{f_{i,j}} + 1 - {p_0}^k_{f_{i,j}} \tag{27}$$

$$w^k_{f_{i,j}} = \frac{L^k_{f_{i,j}}}{\lambda^k_{f_{i,j}}(1 - p^k_{f_{i,j}})} \tag{28}$$

For upper nodes, on the other hand, the queueing delay $w_{f_{i,j}}$ and the packet loss rate $p_{f_{i,j}}$ are obtained almost the same calculation steps, based on the following equations.

$$\lambda_{f_{i,j}} = \frac{\delta_{f_{i,j}}}{m} \tag{29}$$

$$\mu_{f_{i,j}} = \frac{\beta_{f_{i,j}}}{m} \tag{30}$$

$$\rho_{f_{i,j}} = \frac{\lambda_{f_{i,j}}}{\mu_{f_{i,j}}} \tag{31}$$

$$p_{f_{i,j}} = \frac{\rho_{f_{i,j}}^{K_{f_{i,j}}}}{1 + \rho_{f_{i,j}} + \cdots + \rho_{f_{i,j}}^{K_{f_{i,j}}}} \tag{32}$$

$$p_{0\,f_{i,j}} = \begin{cases} \dfrac{1 - \rho_{f_{i,j}}}{1 - \rho_{f_{i,j}}^{K+1}} & (\rho_{f_{i,j}} \neq 1) \\ \dfrac{1}{K+1} & (\rho_{f_{i,j}} = 1) \end{cases} \tag{33}$$

$$L_{q\,f_{i,j}} = \begin{cases} \dfrac{\rho_{f_{i,j}}}{1 - \rho_{f_{i,j}}} - \dfrac{\rho_{f_{i,j}}(K\rho_{f_{i,j}}^{K} + 1)}{1 - \rho_{f_{i,j}}^{K+1}} & (\rho_{f_{i,j}} \neq 1) \\ \dfrac{K(K-1)}{2(K+1)} & (\rho_{f_{i,j}} = 1) \end{cases} \tag{34}$$

$$L_{f_{i,j}} = L_q + 1 - p_0 \tag{35}$$

$$w_{f_{i,j}} = \frac{L_{f_{i,j}}}{\lambda_{f_{i,j}}(1 - p_{f_{i,j}})} \tag{36}$$

2.5.2 Processing Time at Node

For determining the node processing delay, the M/G/1/PS queueing model with r of parallelsim 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.

$$E[R] = \frac{\rho^r}{1 - \rho}\frac{E[S^2]}{2E[S]} + \frac{1 - \rho^r}{1 - \rho}E[S] \tag{37}$$

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

$$\lambda_{n_i} = \frac{x}{m} \tag{38}$$

$$x = \begin{cases} \sum_j \sum_{c_{k,f_{i,j}}} \delta^k_{f_{i,j}} r^k_{f_{i,j}} & \text{if Node } n_i \text{ is EdgeNode or UE} \\ \sum_j \delta_{f_{i,j}} & \text{otherwise} \end{cases} \tag{39}$$

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

3 Numerical Evaluation Results and Discussions

3.1 Evaluation Environment

3.1.1 Network Environment

Figure 3 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. Tables 1 and 2 summarizes the network and server parameters utilized in the evaluation.

3.1.2 Power Consumption Setting

The parameters determing the power consumption characteristics for the UE, network interfaces are presented in Table 2 and 3. The physical servers at the four sites comsume 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 comsume the power constantly regardless of the prosessing traffic rate.

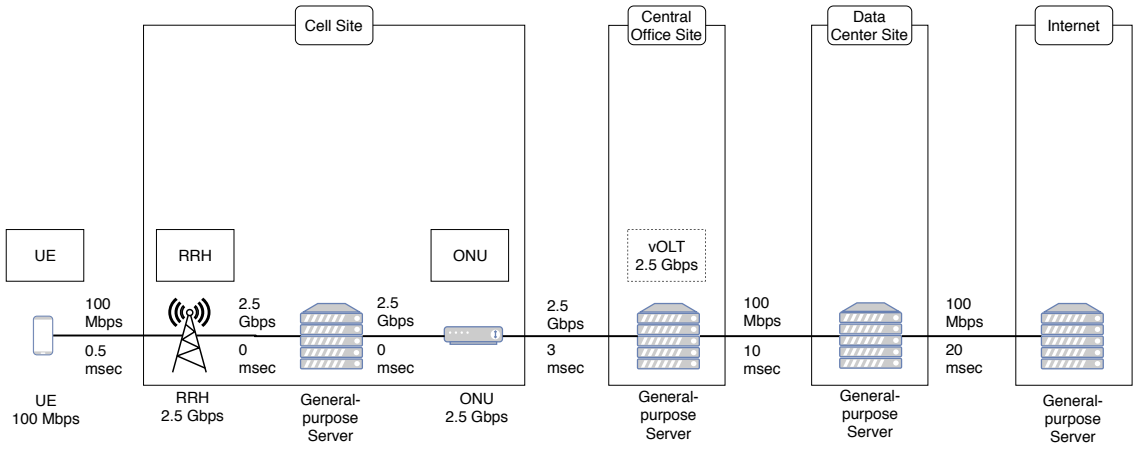


Figure 3: Network environment

Table 1: Link parameter settings

Source node	Destination node	$\tau_{f_{i,x},f_{j,y}}$ [msec]	$\beta_{f_{i,x},f_{j,y}}$ [Gbps]
UE	RRH	0.5	0.1
RRH	ONU	0	2.5
ONU	OLT	3	2.5
OLT	BBU	0	2.5
BBU	EPC	10	0.1
EPC	Application Server	20	0.1

Table 2: Node parameter settings

Node type	π_{n_i} [Gbps]	r_{n_i}	$e_{n_i}^{MAX}$ [W]	$e_{n_i}^{MIN}$ [W]	$e_{n_i}^{SLEEP}$ [W]
UE	0.1	1	0	0	0
RRH	2.5	1	300	150	150
BBU	2.5	1	20	20	20
ONU	2.5	1	50	25	0
OLT	2.5	1	200	100	0
EPC	0.1	1	20	20	20
Application server	0.1	1	20	20	20
General-purpose server			200	200	200

Table 3: Parameter settings of network interfaces

Source Node	Destination Node	$K_{f_{i,j}} - 1$	$e_{f_{i,j}}^{MAX}$ [W]	$e_{f_{i,j}}^{MIN}$ [W]	$e_{f_{i,j}}^{SLEEP}$ [W]
UE	RRH	100	1	0.7	0.01
RRH	UE	100	0.02	0.02	0.02
RRH	ONU	100	4	3	3
RRH	BBU	100	4	3	
ONU	RRH	100	4	3	3
ONU	BBU	100	4	3	3
ONU	OLT	100	10	10	10
OLT	ONU	100	10	10	10
OLT	BBU	100	4	3	3
BBU	RRH	100	4	3	3
BBU	ONU	100	4	3	3
BBU	OLT	100	4	3	3
BBU	EPC	100	4	3	3
EPC	BBU	100	4	3	3
EPC	Application Server	100	4	3	3
Application Server	EPC	100	4	3	3

3.1.3 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 driving. Therefore, the traffic is generated at the regular intervals and it has a strict latency constraint. We set the communicastion 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 3. In the evaluation by using the analysis model in Section 2, we utilze an 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. The configuration of the traffic demands are summarized in Table 4.

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 constaint is located at the site near the UE, we need an additional virtual machine as well as increased power consumption.

Table 4: Application traffic settings

Application No.	Feature	δ_h [Mbps]	T_h [sec]	T_h^{ON} [sec]	σ_h	m [byte]
0	CPRI	2400	1	1	0	1000
1	latency-sensitive	1	1	1	0	1000
		16	1	0.0625	0.9375	1000
		64	1	0.015625	0.984375	1000
2	not tight delay constaint	1–100	1	1	0	1000

3.1.4 Function Placement Patterns

For accommodating the traffic demands in Subsection 3.1.3 on the network in Figure 3, we consider the various placement patterns of network functions (vBBU, vEPC, and application servers) for application 1 and application 2. Figure 4 – 8 depicts the five patterns of the network function placement.

Pettern1: 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 basebend processing are deployed at the central office. (Figure 4)

Pettern2: The application server for application 1 is located at the data center site. (Figure 5)

Pettern3: The application server and vEPC for application 1 are located at the central office site. (Figure 6)

Pettern4: vBBU for baseband processing are deployd at the cell site. The application server and vEPC are also located at the cell site. (Figure 7)

Pettern5: vBBU for baseband processing are deployd at the cell site. The application server and vEPC are located at the central office site. (Figure 8)

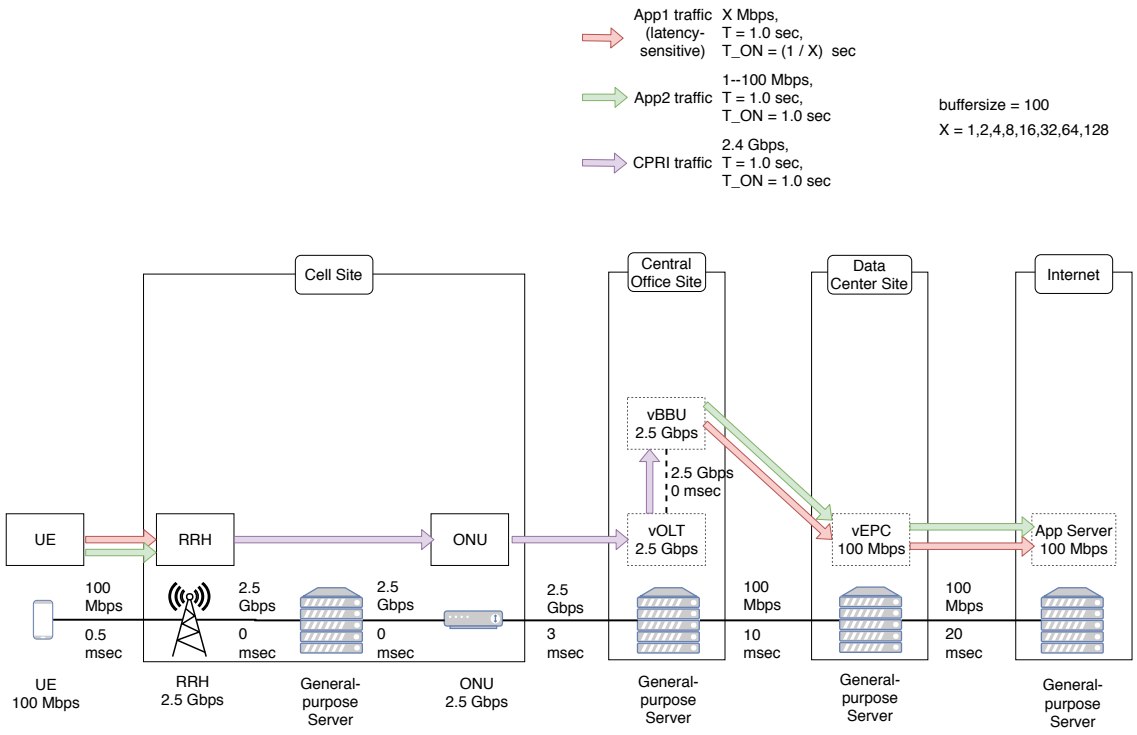


Figure 4: Function placement pattern 1

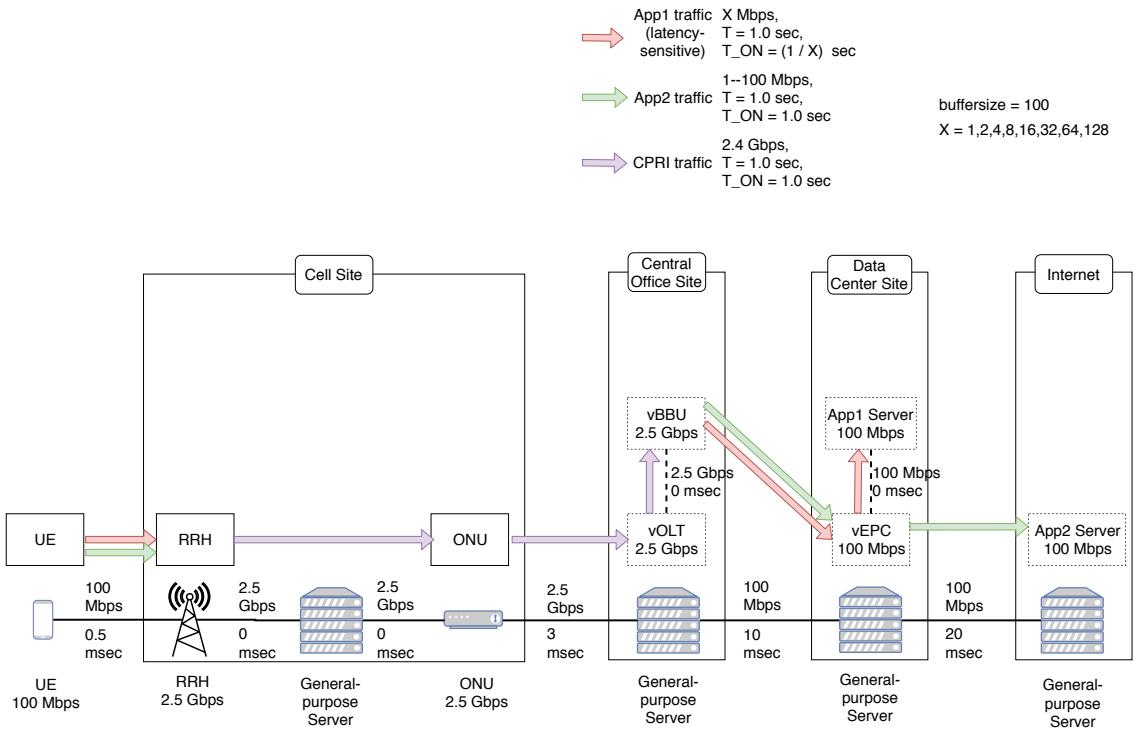


Figure 5: Function placement pattern 2

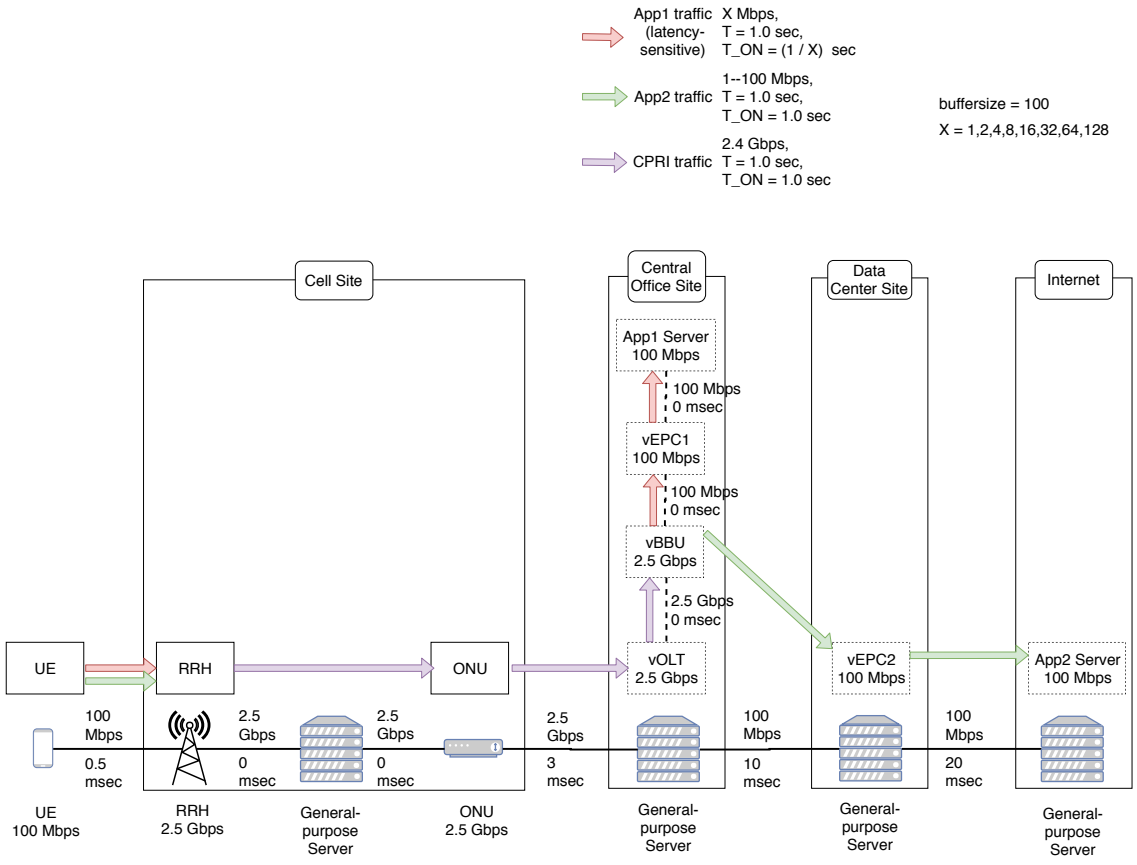


Figure 6: Function placement pattern 3

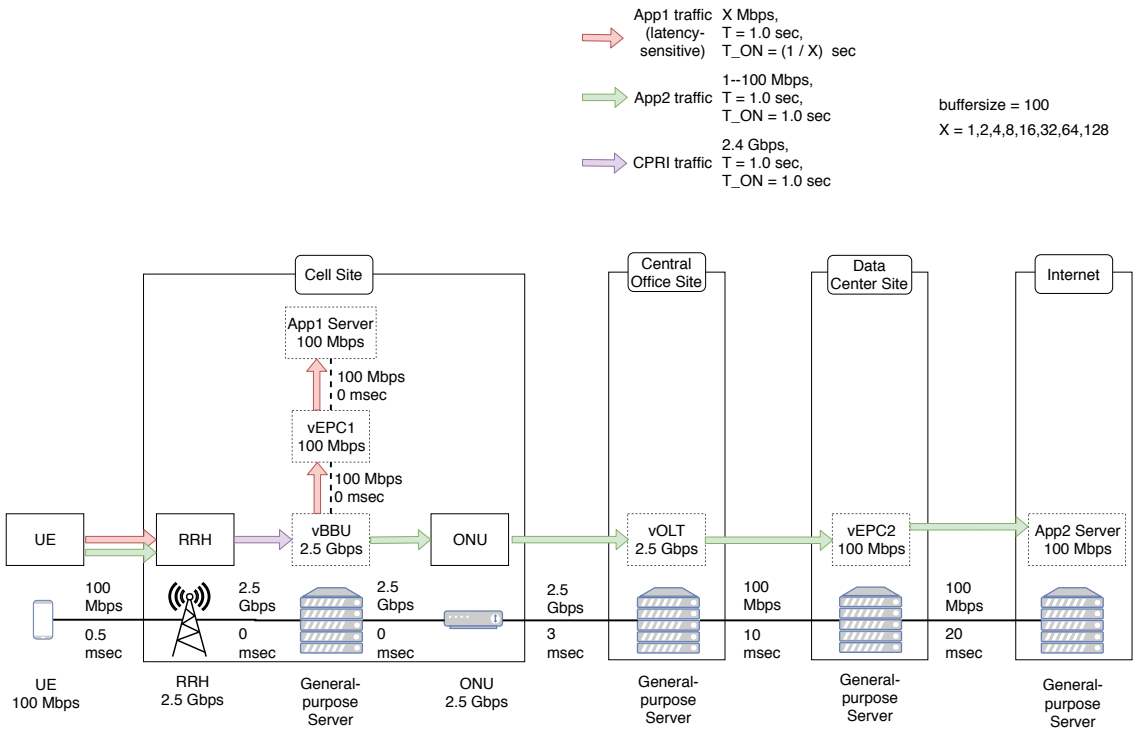


Figure 7: Function placement pattern 4

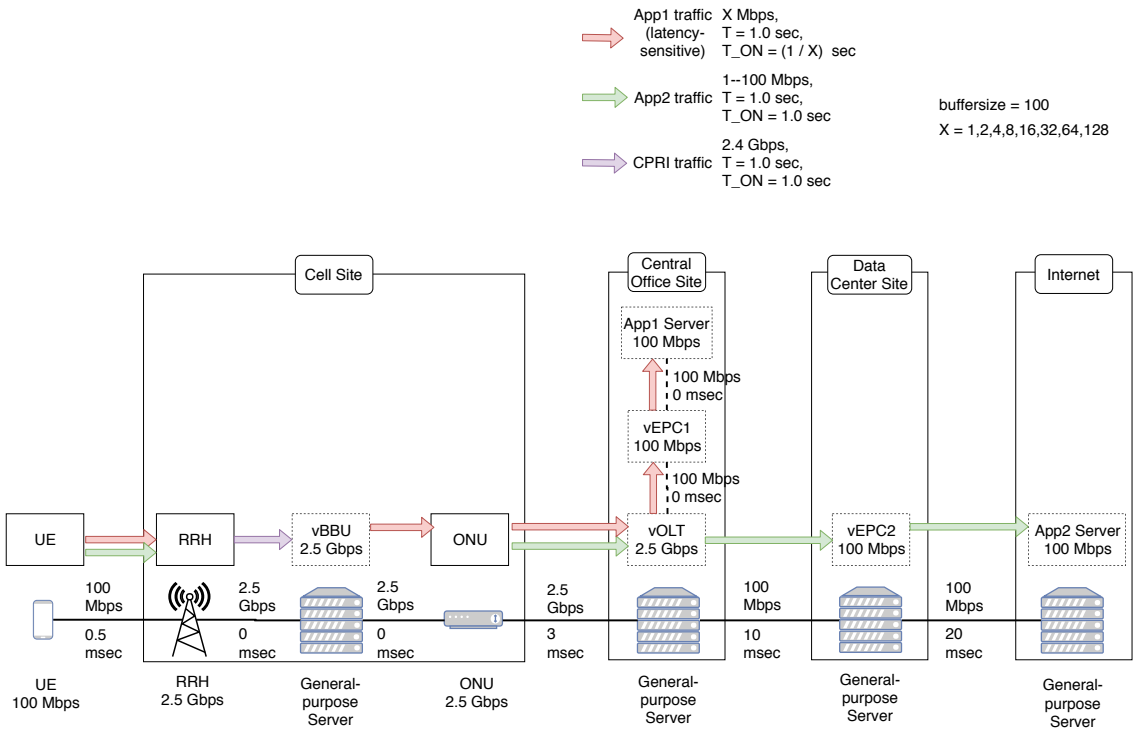


Figure 8: Function placement pattern 5

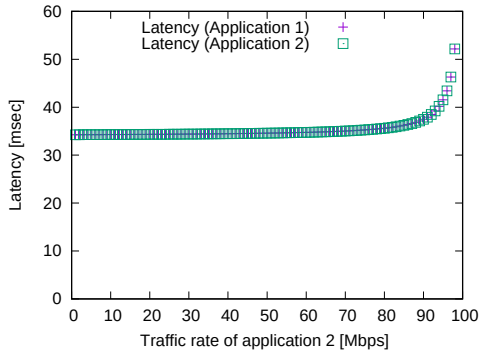
3.2 Results and Discussions

3.2.1 Effect of Traffic Burstiness

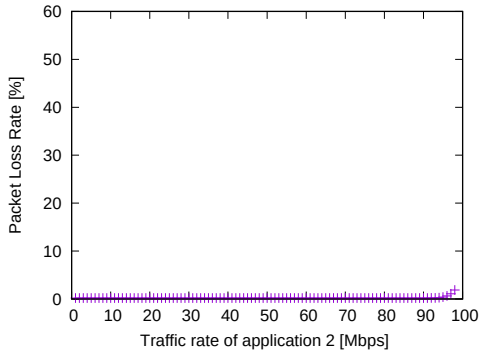
We first evaluate the effect of traffic burstiness on the performance of the application traffic. For the purpose, the communication period of application 1 is set to 1 sec and the data amount transmitted in each ON section is fixed to 128 Kbytes. Then we change the length of ON section to 1 sec, $\frac{1}{2^4}$ sec, $\frac{1}{2^8}$ sec. By decreasing the ON section length the burstiness of the application traffic increases.

Figures 9 – 11 show the evaluation results when we set the ON section length to 1 sec, $\frac{1}{2^4}$ sec, $\frac{1}{2^8}$ sec, respectively. In each figure, we plot the changes in the latency of both applications, packet loss rate and power consumption, as a function of the traffic rate of application 2.

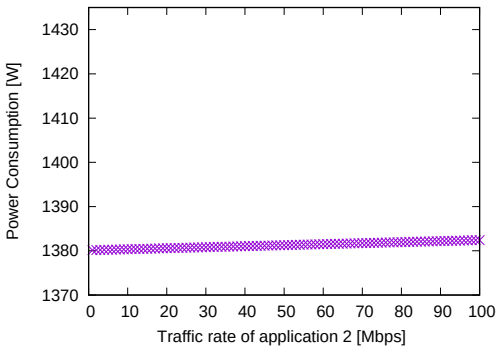
From these figures we can observe that the latency and the power consumption remains unchanged when traffic burstiness changes in application 1. This is because the total traffic transmitted by application 1 does not change. On the other hand, the packet loss characteristics changes significantly. In detail, when we utilze smaller value for the ON section length, the packet loss rate increase earlier with the increase in the traffic rate of application 2. This is because the strong traffic burstiness of application 1 increases the instantaneous load on the network interface, that causes more packet losses.



(a) Latency

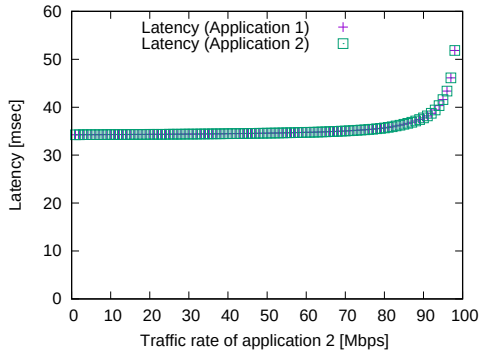


(b) Packet loss rate

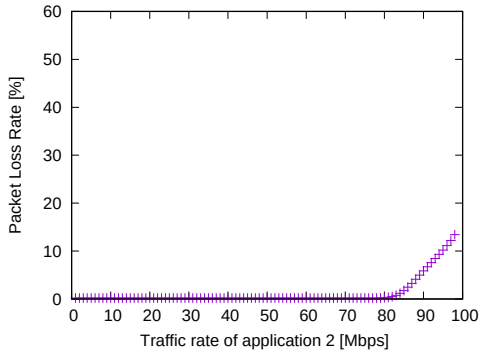


(c) Power consumption

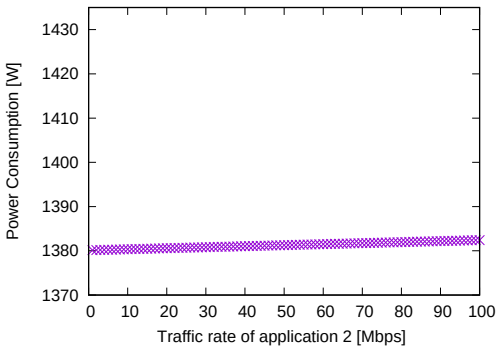
Figure 9: Effect of traffic burstiness: T^{ON} = 1 sec.



(a) Latency

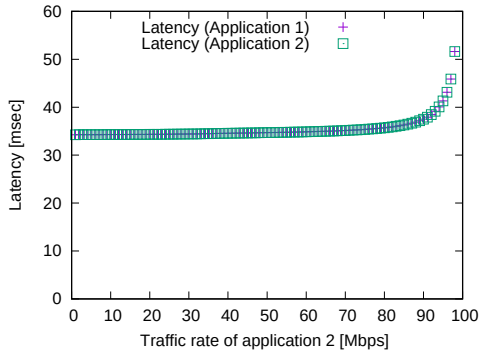


(b) Packet loss rate

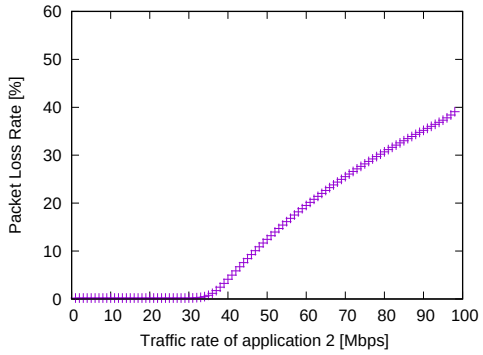


(c) Power consumption

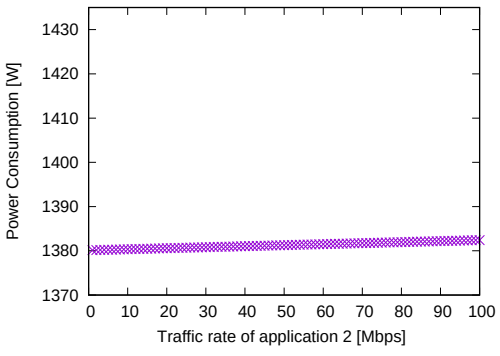
Figure 10: Effect of traffic burstiness: $T^{ON} = \frac{1}{2^4}$ sec.



(a) Latency



(b) Packet loss rate



(c) Power consumption

Figure 11: Effect of traffic burstiness: $T^{ON} = \frac{1}{2^8}$ sec.

3.2.2 Effect of Function Placement

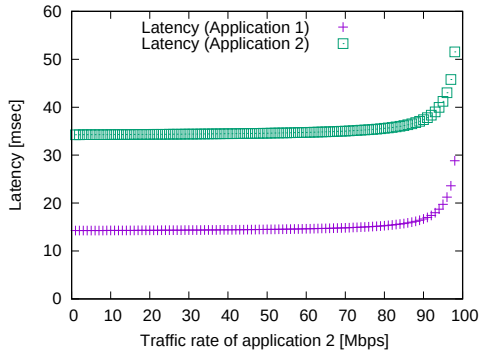
We then assess the effect of function placement for both applications explained in Subsection 3.1.3. As explained in Subsection 3.1.4, we compare five patterns of function placement depicted in Figures 4 – 8. Figures 9, 12 – 15 show the evaluation results of patterns 1– 5, respectively. We set the communiction 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 9 and pattern 2 in Figure 12, 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 addtional 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.

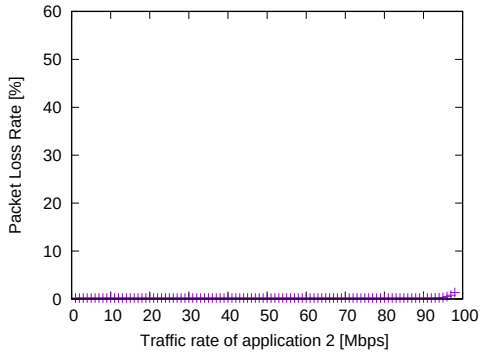
From Figure 13, we can also confirm the same effect by moving servers for application 1 to further nearer location to UEs, where the application server and vEPC for application 1 are located at the central office site. As in the above case, we have decreased latency for application 1 at the sacrifice of increased total power consumption.

In pattern 4, whose evaluation results are presented in Figure 14, 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 patteren 4 can be observed in power consumption found in Figure 14(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.

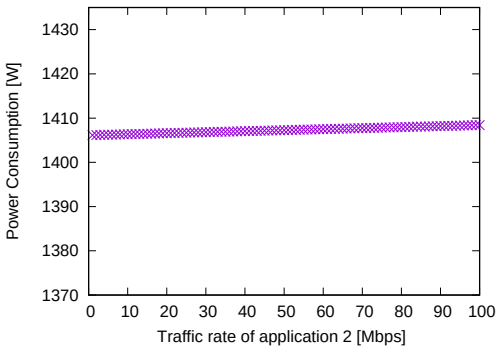
Comparing with pattern 4, pattern 5 in Figure 15 has almost no advantage since the latency of application 1 becomes slightly large by placing the application server and vEPC are located at the central office site. On the other hand the advantage of the small power consumption can be achieved by placing BBU function at the cell site. In the actual situation, however, the power con-



(a) Latency

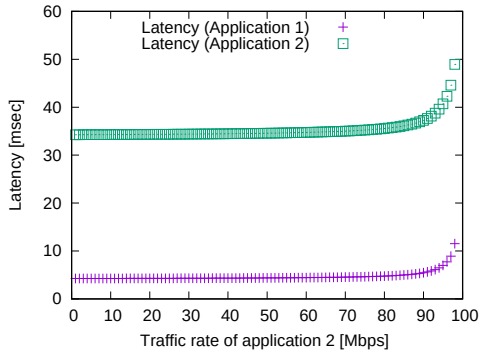


(b) Packet loss rate

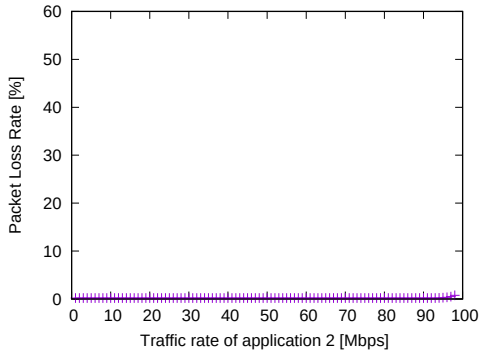


(c) Power consumption

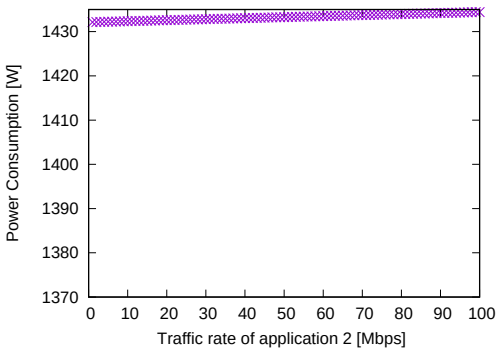
Figure 12: Effect of function placement: Pattern 2



(a) Latency

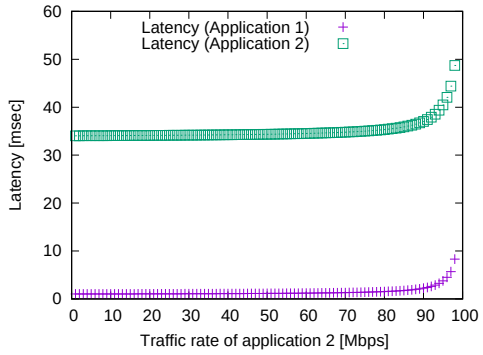


(b) Packet loss rate

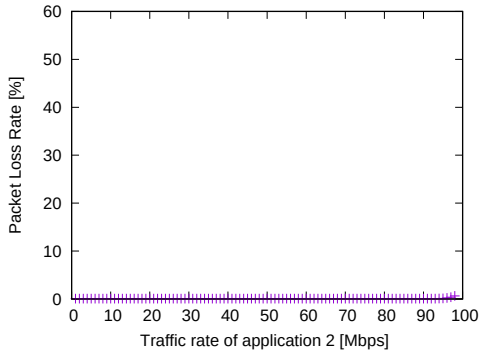


(c) Power consumption

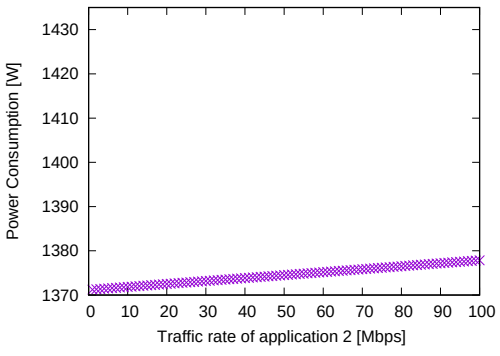
Figure 13: Effect of function placement: Pattern 3



(a) Latency

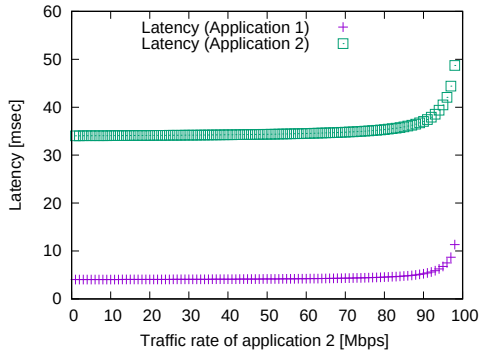


(b) Packet loss rate

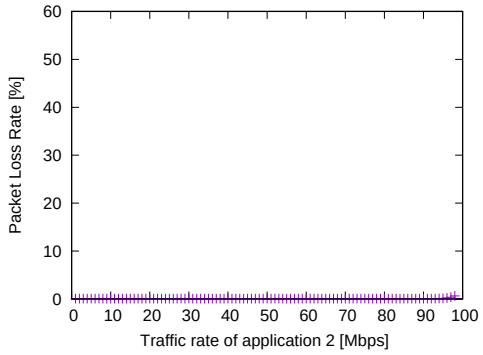


(c) Power consumption

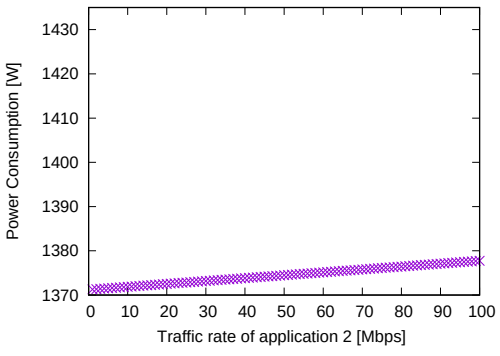
Figure 14: Effect of function placement: Pattern 4



(a) Latency



(b) Packet loss rate



(c) Power consumption

Figure 15: Effect of function placement: Pattern 5

sumption 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.

4 Conclusion

In this thesis, 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. Also, to obtain more reliable numerical results, we need to refine parameter settings of the power consumption characteristics of network interfaces and physical/virtual servers.

Acknowledgment

In promoting this research, many people gave me support. I want to thank Professor Morito Matsuoka. He gave me support in various scenes. In addition, I want to thank Professor Masayuki Murata very much for pointing out from various perspectives about research. Thank you very much for pointing out the points you do not understand about the research you are working on and for giving me the opportunity to learn about research itself and how to approach research. And from day-to-day research to preparation of this thesis, Professor Go Hasegawa gave us very polite guidance in detail. Thanks to him, I was able to continue my research without forgetting my interest in research and purpose, including my attitude towards research and guidelines. I sincerely appreciate it. Thank you very much. And, I thank also students of Matsuoka Laboratory for their support of my laboratory life including support and advice on research when stalled. I heartily thank you.

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] P. Rost, A. Banchs, I. Berberana, M. Breitbach, M. Doll, H. Droste, C. Mannweiler, M. A. Puente, K. Samdanis, and B. Sayadi, "Mobile Network Architecture Evolution Toward 5G," IEEE Communications Magazine, vol. 54, no. 5, pp. 84–91, May 2016.

[3] Z. Pi and F. Khan, "An Introduction to Millimeter-Wave Mobile Broadband Systems," IEEE Communications Magazine, vol. 49, no. 6, pp. 101–107, June 2011.

[4] 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, no. 3, pp. 281–302, Apr. 2003.

[5] C.-L. I, Y. Yuan, J. Huang, S. Ma, C. Cui, and R. Duan, "Rethink Fronthaul for Soft RAN," Communications Magazine, IEEE, vol. 53, no. 9, pp. 82–88, Sept. 2015.

[6] 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.

[7] C. Chang, R. Schiavi, N. Nikaein, T. Spyropoulos, and C. Bonnet, "Impact of Packetization and Functional Split on C-RAN Fronthaul Performance," in *Proceedings of 2016 IEEE International Conference on Communications (ICC)*, pp. 1–7, May 2016.

[8] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network Slicing in 5G: Survey and Challenges," IEEE Communications Magazine, vol. 55, no. 5, pp. 94–100, May 2017.

[9] H. Zhang, N. Liu, X. Chu, K. Long, A. Aghvami, and V. C. M. Leung, "Network Slicing Based 5G and Future Mobile Networks: Mobility, Resource Management, and Challenges," IEEE Communications Magazine, vol. 55, no. 8, pp. 138–145, Aug. 2017.

[10] 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, vol. 22, no. 5, pp. 32–40, Oct. 2015.

[11] X. Costa-Perez, A. Garcia-Saavedra, X. Li, T. Deiss, A. de la Oliva, A. di Giglio, P. Iovanna, and A. Moored, "5G-Crosshaul: An SDN/NFV Integrated Fronthaul/Backhaul Transport Network Architecture," IEEE Wireless Communications, vol. 24, no. 1, pp. 38–45, Feb. 2017.

[12] D. Kreutz, F. M. V. Ramos, P. E. Veríssimo, C. E. Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-Defined Networking: A Comprehensive Survey," Proceedings of the IEEE, vol. 103, no. 1, pp. 14–76, Jan. 2015.

[13] B. A. A. Nunes, M. Mendonca, X. Nguyen, K. Obraczka, and T. Turletti, "A Survey of Software-Defined Networking: Past, Present, and Future of Programmable Networks," IEEE Communications Surveys Tutorials, vol. 16, no. 3, pp. 1617–1634, Thirdquater 2014.

[14] 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, no. 1, pp. 236–262, Firstquarter 2016.

[15] B. Han, V. Gopalakrishnan, L. Ji, and S. Lee, "Network Function Virtualization: Challenges and Opportunities for Innovations," IEEE Communications Magazine, vol. 53, no. 2, pp. 90–97, Feb. 2015.

[16] P. Mahadevan, P. Sharma, S. Banerjee, and P. Ranganathan, *A Power Benchmarking Framework for Network Devices*, pp. 795–808. Springer Berlin Heidelberg, May 2009.

[17] D. Gross, J. F. Shortle, J. M. Thompson, and C. M. Harris, *Fundamentals of Queueing Theory, 4th Edition*. John Wiley and Sons, Inc., 2008.