Minimum Delay Function Placement and Resource Allocation for Open RAN (O-RAN) 5G Networks

Nasim Kazemifard and Vahid Shah-Mansouri
School of Electrical and Computer Engineering, College of Engineering,
University of Tehran, Tehran 14395-515, Iran
emails: {na.kazemifard@gmail.com, vmansouri@ut.ac.ir}

Abstract

Digitalization is a journey that has been started and put ICT industry in a crucial situation to provide required infrastructure for diverse range of data hungry, short tempered applications and services. One of the main technologies that will pave the way towards new digital ecosystem is fifth Generation of mobile technology. To meet 5G network service requirements, innovative architectures, technologies and standards focusing on cloudification are employed. Cloudification of network functions along with the use of virtualized network functions (VNFs) and containerized network functions (CNFs) allows agile and scalable service provisioning. The use of VNFs and CNFs has been started from core and networking middleboxes but then extended to RAN functions. Open radio access network (O-RAN) proposes an interoperable and standard architecture for cloudified RAN. The main idea behind this architecture is to make RAN more flexible. O-RAN allows different layers of RAN to split and deployed as virtual function and openly communicate with each other for service provisioning. In this paper, we model an end to end mobile network operator (MNO) employing O-RAN. We consider a mobile network architecture, with three layer hierarchical data centers (Local, Regional, and Core) to add flexibility in resource allocation, and increase reliability, taking the advantages of O-RAN. MNO receives various service function requests (SFRs) requiring accommodation on the network. We assume RAN and core functions are deployed as CNFs on the data centers. Users of SFRs connects to remote radio heads

(RRH) to receive the service. In this paper, we mathematically model the CNF placement and resource allocation of an O-RAN enabled LTE/5G network while trying to minimize the end to end delay of the data plane. We study the problem in two different cases. First, we assume that the SFR traffic traverse through a single path across the RAN functions and model this problem. This is a mixed integer non-linear programming problem. With some change of variables, we make it a linear mixed integer programming problem but it is still non-trivial to solve. Then, we model the problem for the case where traffic of an SFR can be split and be served via multiple CNFs. We proposed a gradient based scheme to solve the minimum delay problem in this case.

Experimental results indicate that by increasing the number of service requests in a network, the proposed GBMD(Gradient-Based Minimum Delay algorithm serves up to 90% e2e Delay decrease. Another improvement on the performance of a network will occur by levering GBMD algorithm for around 72% e2e Delay reduction in case of limited resources.

Keywords: Open Radio Access Network (O-RAN), Hierarchical Data Center, Containerized Network Function (CNF), Resource Allocation, Minimum Delay Function Placement.

1. Introduction

Day to day life is being affected by digital ecosystem. Increasing in variety of services and applications brings more demand from customers in any aspect of life, such as shopping, learning, health care, traveling, sport, etc. Besides application developers and OTT players, Telco Operators play a key role on realization of this life style. Operators, mainly MNOs, in the role of infrastructure and service providers should cope with such increasing demand from customers and application developers, in terms of huge investment, agile strategic plans, architectural review, business review, etc. [1]. In other words, besides strategic and business changes, they have to think about technical conversions in their networks.

Taking a look over standardization trends and technology roadmap indicates an inevitable migration for telecom operators toward virtualization. There is a history behind current mobile network technologies. Starting from first generation, to 2G with poor data throughput in form of GPRS, continued with third generation of mobile network (3G) that supports better speed of data up to 8 Mbps. More data speed requirements lead standardization bodies to changing priority in architecture based on data which happened in 4G or LTE solution with 150 Mbps. In this generation, many advanced features are added to increase data throughput such as carrier aggregations, massive MIMO, and etc. to reach up to 1Gbps data rate per customer[2].

In legacy mobile network architecture, shown in Fig.1, there are some physical functions that serve customers with voice and data services. The network consists of radio access network, which connects customer to Antennas, and Backhaul part plus IP backbone (IPBB) network that includes high speed switches and routers to deliver user data to core part of the network. The functionality of serving customer data will happen in several Physical Network Functions called PNF, such as SGW, MME, HSS, etc.

This fixed and physical oriented architecture can not support all ambitious targets for future solutions. Customers need more data, in very dense areas, with a minimum E2E delay and reliability. Fifth generation of network is planned to meet these targets. 5G will bring more flexibility and reliability to network providers via NR (New Radio) [3], and 5GC (5G Core). The minimum required bandwidth and new products are not the concerns in this paper. But, there are some enablers that will use for migration towards 5G. One of these enablers is Virtualization of network functions which is the fundamental of Telco data centers and orchestration in network that we will use for our work.

In order to reduce CAPEX and OPEX, minimize time-to market of new services, and maintain profit by creating new revenue streams, Operators are adopting network function virtualization (NFV) technologies and are shifting from physical hardware towards NFV platforms with help of softwarization and

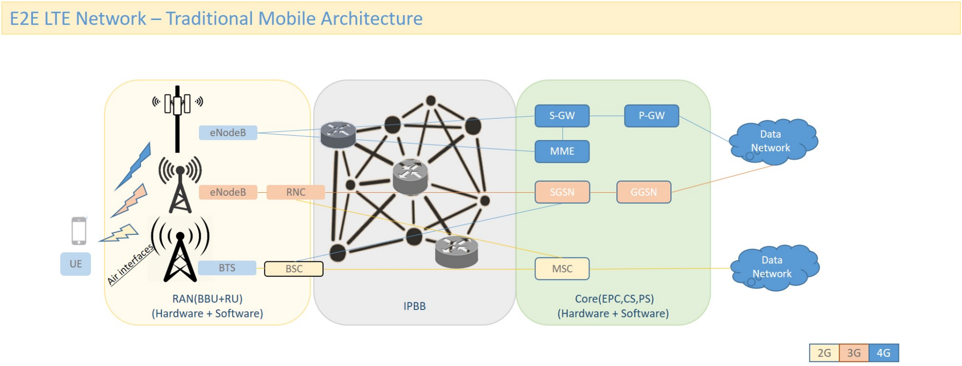


Figure 1: E2E 2G/3G/4G Legacy Network

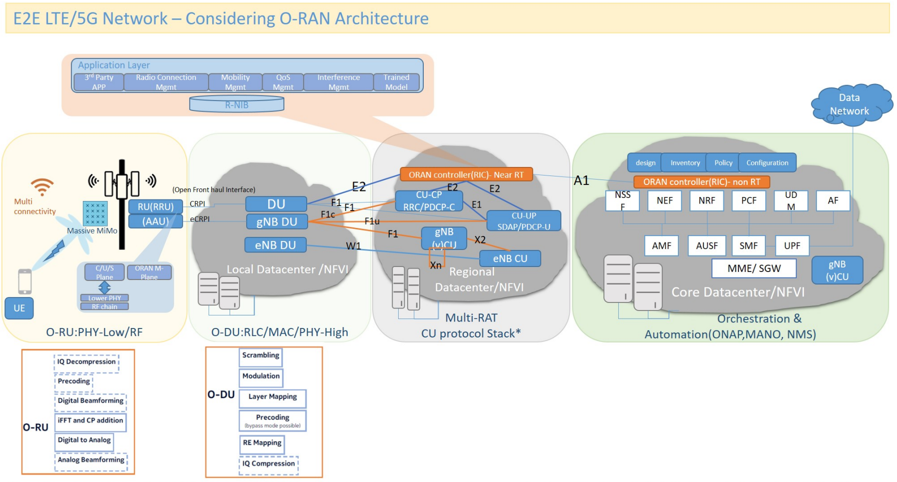
cloudification [4]. Through NFV, network functions become virtualized and are called Virtualized Network Functions (VNFs) in standards. VNFs are deployed as VMs (virtual machines) with hypervisors such as Linux KVM or VMware vSphere on Commercially Off-The-Shelf hardware (COTS) and bring vendor independency for operators. [5]For light weight implementation of functions, in some cases, Containerized Network Functions (CNFs) are employed instead.Transitioning from a hardware-centric to a software-centric paradigm is challenging. Different targets for 5th generation of mobile technology defines tough targets such as five nine reliability, one millisecond latency, and up to 10Gbps throughput [6]. 5G networks simultaneously support several services with different requirements in categories including Enhanced Mobile Broadband (eMBB), Ultra Reliable Low Latency Communications (URLLC), and Massive Machine Type Communications (mMTC) [7]. These technologies lead to redesign the service provisioning architecture from dedicated hardware appliances or middleboxes to white-boxes. By moving packet-processing activities from proprietary hardware middleboxes to virtualized entities, dependency on underlying hardware and vendors, which generally push operators into Vendor Lock-In situation, will be reduced. To handle VNF/CNF functions, end-to-end cloud based architecture should be employed. [8]

Hierarchical data centers are used in operator networks to distribute the deployment of the functions. VNFs or CNFs can be deployed in local, regional, and core data centers. Each service contains a service chain composing of several VNFs/CNFs connected back to back. Various VNFs/CNFs of a service can be placed in different data centers.

Open RAN (O-RAN) is a vendor-neutral disaggregation of RAN at both the hardware and software levels on general purpose processor-based platforms which implements an open interface between components RU/CU/DU using hardware- and software-defined functions, and brings cloud-scale economics and agility to radio access part of the network by means of modular softwarization for capacity management, increasing reliability and availability, easily and quickly network tuning with scale-up/scale-down designs rather than expensive hardware expansion in vendor locked situation. Admitting new services and applications, realization of network slicing, and DevOps concept is a result of implementing O-RAN in network. In comparison to traditional RAN which had integrated RRU and BBU with high cost last mile transport network, deployment and management in O-RAN is flexible according to its agnostics front haul. Programmable application and network adaptability in O-RAN can be compared to pre-programmed and fixed control logic and fixed network resources in legacy network as well.

However, O-RAN splits the control-plane (CP) from the user-plane (UP) through E1 interface (3GPP standard) inherited from SDN concept, and will handle Radio Resource Management functions in terms of NG-RRM via hierarchical (Non-RT and Near-RT) RAN Intelligent Controller (RIC) with A1 and E2 interfaces as shown in fig.2.

In the next section of the paper, we consider a system model by combining NFV and O-RAN reference models. A three layer hierarchical cloud based network architecture is used for NFV deployment. Mobile network operator receives service function requests (SFRs) which have different resource requirements and place them on the virtual infrastructure. We model the problem of CNF placement and resource allocation of various SFRs for RAN and core functions while minimizing the end to end delay for data plane. The system model is shown in



*The existing interface definitions for F1/W1/E1/X2/Xn provided by 3GPP, will be enhanced to support interoperation among multivendors and the CU provided by TEMs offer a regional CP and UP anchor for DUs.

Figure 2: E2E LTE/5G Network- Considering O-RAN Architecture, with 3 layer Data Centers

Fig. 2. The contributions of the paper are listed as follows:

- We mathematically formulate the problem of resource allocation and CNF placement on a virtualized 5G network employing O-RAN.
- First, we consider the case where the traffic of an SFR traverses via a single path through CNFs of its chain. It is a nonlinear mixed integer programming problem. Via some re-formulations, we convert it to a linear programming problem. However, it is still non-trivial.
- We then formulate the case where SFR traffic can traverse multiple pathes. We propose a gradient based solution which achieves the optimal solution efficiently.
- We evaluate and compare the performance of the proposed schemes via simulations.

The rest of the paper is organized as below. After presenting the state of the art, the system model is introduced in Section 3. The single path placement and resource allocation problem are modeled in Section 4. In Section 5, we propose the alternative multiple path problem and the gradient based efficient solution. Finally, investigation of the performance of optimal solution with exhaustive approach via some simulations is shown in Section 6.

2. State of the art

The European Telecommunications Standards Institute (ETSI) [9] proposed an architectural framework for NFV. The resources of NFV Infrastructure (NFVI) are compute (i.e., processing), storage, and network capacity. A Virtualized Infrastructure Manager (VIM) (e.g., OpenStack [10] or Kubernetes) manages NFVI and utilizes existing virtualization technologies to provide virtual resources for VNFs and CNFs. The CNFs represent virtualized instances of different network functions [11]. In this reference model, there is an Orchestrator (i.e., MANO) that manages the life-cycle of CNFs of the service. It utilizes resource allocation and placement algorithms to ensure optimal usage of both physical and software resources [11].

Cloudification in core network is employed and implemented in many operators networks until now. Nevertheless, in radio access network (RAN), recently, standardization is completed and new reference models are proposed. Previously, C-RAN architecture was proposed and its standardization is introduced in [12], [13], [14]. There is an study in which the C-RAN architecture is proposed using available infrastructure such as site locations and transmission links, with the aim of cost reduction [15]. In, [16] also an investigation on a cross-layer resource allocation model for C-RAN to minimize overall system power consumption is proposed. In another study, energy minimization and resource allocation based on C-RAN architecture is modeled and described [17]. Afterwards, O-RAN alliance defines a cloud based architecture evolving towards an open, agile, flexible, and programmable RAN paradigm [18], [19]. In 2013, M. Yang, *et. al.* proposed an architecture for software-defined RAN based on virtualization [20], in which three main parts, wireless spectrum resource pool (WSRP) that enables virtual RRUs (vRRUs) to support different wireless protocols in one shared pRRU, Cloud Computing Resource Pool (CCRP) that includes a large amount of physical processors to support a high speed cloud computing network via virtualized functions such as vBBUs and vBSCs, and an SDN controller to play a role of control plane to execute the strategies of each vBBU and vBSC that contains a SDN agent to communicate with controller through SDN protocol. Although the proposed architecture is considerable in access part of the network, it did not take care of Core functions, and there is no End to End view of O-RAN implementation. O-RAN alliance has proposed O-RAN Controller instead of SDN controller in this architecture which has more complicated functionality and we use it in this paper. Another Cloud-RAN architecture is presented in [21], that supports indoor front-ends and the BBU server can be installed inside the building that is covered with multiple RRH considering DAS algorithms instead of TAS. In the proposed architecture, controllers and MANO roles are not considered. In 2019, an application on O-RAN is disclosed [22] that visualizes an architecture for a physical layer functional split between the CU and the RUs to maximize the efficiency of the transport and increase the flexibility of supporting required features for virtualization and commercialization of cloud RAN technology. Although this patent is limited only to cloud-RAN splitting aspects; but from implementation point of view, it proposes a valuable methodology. In Reference [23], an approach from C-RAN to O-RAN implementation is proposed, and it is mentioning that by developing embedded intelligence, the O-RAN architecture aims to not only extend the SDN concept of decoupling the control-plane (CP) from the user-plane (UP), but also to enhance the traditional RRM functions with embedded intelligence via RAN Intelligent Controller (RIC) near-RT..

3. System Model

We consider a virtualized multi-layer cellular network with several hierarchical data centers including local (i.e., edge), regional (i.e., fog), and core data centers. Data center infrastructure allows support of network function virtualization (NFV) which not only gives more flexibility to network designer and orchestrator, but also increases reliability and availability in terms of resource distribution and utilization [24]. Local data centers are those close to radio sites with limited amount of resources. Each local data center can connect to several regional data centers physically via high capacity fiber optical links. The regional data centers have higher capacity and amount of resources; these nodes receive data from local data centers and send to core data center. Core data centers are limited in terms of number but they are significantly rich in terms of resources. Core data centers supports running virtualized EPC and 5GC cores [6]. Multiple instances of core can be employed to increase reliability and availability.

We consider a cellular network employing O-RAN architecture. Fig.2 shows the system model. In O-RAN architecture, low physical layer part (i.e., open remote unit (O-RU)) is running on RRH. It includes several physical layer tasks such as IQ decompression, precoding, digital beamforming, iFFT CP addition, and D2A. Using open franthaul interfaces such as CPRI in LTE and eCPRI in 5G, O-RU connects to RAN distributed unit (DU). Virtualized high physical layer, MAC, and RLC functions run on DU which is deployed on local data centers as a single containerized network function (CNF). We call such function as DU-CNF. The upper layer functions of the RAN in 5G includes RRC, PDCP-C, PDCP-U, and SDAP which run as one CNF on the regional data centers. We call this function CU-CNF. O-RAN Intelligent controllers (RIC) near-RealTime(RT) are also deployed in regional data center to be close to CU-CNF functions and are connected to them via an interface called E2. Cellular core functions runs on core data centers. Slicing is used to isolate different services and reach the required SLA levels [25].

Customers receive service by connecting to the RRHs which are connecting to RAN part, and requesting for services by attaching to RRHs. In order to guarantee Quality of Service (QoS) in Cloud based architecture, QoS parameters such as throughput, jitter, and delay should be rephrased to resource allocation in terms of capacity, storage and processing capability [26], [18].

It is important to notice that there is a data base called R-NIB in RIC Near-RT, and some application layer functions are handled in it, such as 3^{rd} party, Radio Connection, Mobility, QoS, Interface managements. RIC near-RT in Regional data centers connects logically via A1 interface to RIC near-RT in core data centers.Orchestration and automation in terms of ONAP, MANO, and NMS are the functions added to NG Core rather than traditional core network.

There are different services in the network requiring end to end resources from the system including RRH, DU-CNF, CU-CNF, and core network resources. We call them service function request (SFR). Let $\mathcal{I}$ denote the set of SFRs. Each SFR has a service function chain which consists of a DU-CNF placed on the local data center, a CU-CNF placed on regional data center, and a Core-CNF placed on the core data centers. We assume that an SFR has a transmission rate on its data plane. For the sake of simplicity, we assume that this rate is constant throughout the system in RRH, DU-CNF, CU-CNF, and core. Let θ_i denote the service rate of SFR $i \in \mathcal{I}$. θ_i is an accumulative rate for all users of the i^{th} SFR. Table 3 includes list of variables.

4. Optimal Placement of CNFs for Single Path Service Traffic

In this section, we model the problem of RAN and core CNF placement and resource allocation assuming that traffic flow of an SFR traverse a single path across the functions. This means, only a single DU-CNF, a single CU-CNF, and a single Core-CNF provide service for each SFR. We shall notice that to support isolation via slicing, ??????? We first study the constraints of the problem. Let $\mathcal{A}$ denote the set of local data center servers. We use boolean variable α_i^a to

Table 1: Simulation Parameters

Parameter name	Value
$\mathcal{I}$	Set of SFRs
$\mathcal{A}$	Set of servers of local data center
$\mathcal{B}$	Set of servers of regional data centers
$\mathcal{C}$	Set of servers of core data centers
α_i^a	decision binary variable indicating if SFR i using server $a \in \mathcal{A}$
β_i^b	decision binary variable indicating if SFR i using server $b \in \mathcal{B}$
γ_i^c	decision binary variable indicating if SFR i using server $c \in \mathcal{C}$
y_i^L, y_i^R, y_i^C	processing requirement of SFR i on local, regional, and core data centers.
z_i^L, z_i^R, z_i^C	storage requirement of SFR i on local, regional, and core data centers.
$\rho_i^L, \rho_i^R, \rho_i^C$	processing capacity of local, regional, and core data centers.
S_i^L, S_i^R, S_i^C	storage limit of SFR i on local, regional, and core data centers.
d_a^L, d_b^R, d_c^C	rate capacity of local, regional, and core data centers server.
θ_i	data rate requirement of SFR i

denote if DU-CNF of the i^{th} SFR is deployed on the a^{th} server of local data center. Let y_i^L and z_i^L denote respectively the capacity and storage requirement of SFR i from the local center. This is indeed the requirement of DU-CNF of SFR i. The resource constraints of the data centers mandate that

$$\begin{aligned} \textstyle\sum_{i\in\mathcal{I}} \alpha_i^a y_i^L &\leq \rho_a^L, \quad \forall a \in \mathcal{A}, \\ \textstyle\sum_{i\in\mathcal{I}} \alpha_i^a z_i^L &\leq S_a^L \quad \forall a \in \mathcal{A}, \\ \textstyle\sum_{i\in\mathcal{I}} \alpha_i^a \theta_i &\leq d_a^L, \quad \forall a \in \mathcal{A}, \end{aligned} \tag{1}$$

where ρ_a^L, S_a^L, and d_a^L denote the total capacity, storage, and transmission rate of the a^{th} local data center.

The same constraint can be used for regional data centers. Let $\mathcal{B}$ denote the set of regional data center servers. Let boolean variable β_i^b denote if CU-CNF of the i^{th} SFR is deployed on the b^{th} server of regional data center. CU-CNF of the i^{th} SFR has capacity and storage requirement form the regional data center denoted respectively by y_i^R and z_i^R. The resource constraints for servers of the regional data center mandates that

$$\begin{aligned}
\textstyle\sum_{i\in\mathcal{I}} \beta_i^b y_i^R &\leq \rho_b^R, \quad \forall b \in \mathcal{B}, \\
\textstyle\sum_{i\in\mathcal{I}} \beta_i^b z_i^R &\leq S_b^R, \quad \forall b \in \mathcal{B}, \\
\textstyle\sum_{i\in\mathcal{I}} \beta_i^b \theta_i &\leq d_b^R, \quad \forall b \in \mathcal{B},
\end{aligned} \tag{2}$$

where ρ_b^R, S_b^R, and d_b^R denote the total capacity, storage, and transmission rate of the b^{th} server of regional data center.

Similarly, we have constraints for core data centers. Let $\mathcal{C}$ denote the set of core data center servers and boolean variable γ_i^c denote if the core CNF of the i^{th} SFR is deployed on the c^{th} server of core data center. Core CNF of the i^{th} SFR has capacity and storage requirements from the core data centers denoted respectively by y_i^C, z_i^C. The resource constraints of the core data center mandates that

$$\begin{aligned}
\textstyle\sum_{i\in\mathcal{I}} \gamma_i^c y_i^C &\leq \rho_c^C, \quad \forall c \in \mathcal{C} \\
\textstyle\sum_{i\in\mathcal{I}} \gamma_i^c z_i^C &\leq S_c^C \quad \forall c \in \mathcal{C} \\
\textstyle\sum_{i\in\mathcal{I}} \gamma_i^c \theta_i &\leq d_c^C, \quad \forall c \in \mathcal{C}
\end{aligned} \tag{3}$$

where ρ_c^C, S_c^C, and d_c^C denote the capacity, storage, and transmission rate of the c^{th} core data center. We notice that SFR i at each data center level uses one CNF on one server. Therefore, we have

$$\begin{aligned}
\sum_{a\in\mathcal{A}} \alpha_i^a &= 1, \quad \forall i \in \mathcal{I}, \\
\sum_{b\in\mathcal{B}} \beta_i^b &= 1, \quad \forall i \in \mathcal{I}, \\
\sum_{c\in\mathcal{C}} \gamma_i^c &= 1. \quad \forall i \in \mathcal{I}.
\end{aligned} \tag{4}$$

The required service rate of SFR i, θ_i, can be provided by different RRHs. Let θ_i^k denote the rate of SFR i served by RRH k. Let $\mathcal{K}$ denote the set of RRHs.

Le Γ_k denote the transmission limit of the kth RRH. The RRH rate constraint mandates that

$$\sum_{i\in\mathcal{I}} \theta_i^k \leq \Gamma_k, \quad \forall k \in \mathcal{K}. \tag{5}$$

The processing delay is defined as the delay imposed at the CNF for processing of the packets of an SFR. Assuming an M/M/1 model for the processing delay [27], the processing delay of an CNF is equal to $1/(\psi - \varphi)$ where ϕ is the service rate of the CNF and φ is the arrival rate of packets to the CNF. Different CNFs on a server share the processing resources of the server. We remind that ρ_a denote the processing capacity of server a in local data center. Let F^L be a constant where $\rho_a F^L$ denotes the data rate at which a server $a \in \mathcal{A}$ can process the packets at local data center server. This is considered as the service rate of server a. Since we assume CNFs co-exist on a server, the processing resources of the server are shared between the CNFs running on that. The total arrival rate of server a is $\sum_{i\in\mathcal{I}} \alpha_i^a \theta_i$. Delay of DU-CNF for SFR $i \in \mathcal{I}$ is

$$D_i^{\text{DU}} = \frac{1}{\rho_a^L \times F^L - \sum_{i\in\mathcal{I}} \alpha_i^a \theta_i}, \quad \text{if } \alpha_i^a = 1. \tag{6}$$

For regional and core data centers, the processing delay of CU-CNF, D_i^{CU}, and core CNF, D_i^{Core}, are similarly as follows:

$$D_i^{\text{CU}} = \frac{1}{\rho_b^R \times F^R - \sum_{i\in\mathcal{I}} \beta_i^b \theta_i}, \quad \text{if } \beta_i^b = 1, \tag{7}$$

$$D_i^{\text{Core}} = \frac{1}{\rho_c^C \times F^C - \sum_{i\in\mathcal{I}} \gamma_i^c \theta_i}, \quad \text{if } \gamma_i^c = 1. \tag{8}$$

Next, we consider the transmission delay between the data centers. We consider a constant transmission costs between the data centers. Let $\varepsilon^{L->R}(a, b)$ and $\varepsilon^{R->C}(b, c)$ denote the cost of using links from local data center server a to regional data center server b and from regional data center server b to core data center server c, respectively. The transmission cost for SFR i is non-zero between local server a and regional server b if both α_i^a and β_i^b is non-zero. We introduce binary variable $\varpi_{a,b}^i$ which is one if both variables α_i^a and β_i^b are one. We can relate $\varpi_{a,b}^i$ to α_i^a and β_i^b as

$$1 - \varpi_{a,b}^i \leq (1 - \alpha_i^a) + (1 - \beta_i^b), \quad \forall i \in \mathcal{I}, a \in \mathcal{A}, b \in \mathcal{B}. \tag{9}$$

Equation (9) guarantees that $\varpi^i_{a,b}$ is one if α^a_i and β^b_i are one. If one of α^a_i and β^b_i are zero, then $\varpi^i_{a,b}$ can be one and zero. Since we are minimizing the delay, $\varpi^i_{a,b}$ takes value 0 in case it has option between one and zero. Considering the transmission delay, the end-to-end processing and transmission delay can be written as

$$\begin{aligned} D_i &= \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} \varpi^i_{a,b}\varepsilon^{L->R}(a,b)\theta_i \\ &+ \sum_{b\in\mathcal{B}}\sum_{c\in\mathcal{C}} \varpi^i_{b,c}\varepsilon^{R->C}(b,c)\theta_i \\ &+ \frac{\alpha^a_i}{\rho^L_a \times F^L - \sum_{i\in\mathcal{I}} \alpha^a_i\theta_i} + \frac{\beta^b_i}{\rho^R_b \times F^R - \sum_{i=\gamma^c_i}^{I} \beta^b_i\theta_i} \\ &\quad + \frac{\gamma^c_i}{\rho^C_c \times F^C - \sum_{i\in\mathcal{I}} \gamma^c_i\theta_i} \end{aligned}$$

The minimum delay problem can be formulated as

$$\begin{aligned} \min \qquad & D_{\text{total}} = \sum_{i\in\mathcal{I}} D_i \\ \text{subject to} \quad & (1), (2), (3), (4), (5). \end{aligned} \tag{10}$$

This is a binary non-convex optimization problem which is known to be non-trivial to solve optimally. We perform a series of changes to make this problem a linear binary programming problem. For equations in (6)-(8), we use inequality instead of equality. Then, we have

$$\min \sum_{i\in\mathcal{I}}\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} \varpi^i_{a,b}\varepsilon^{L->R}(a,b)\theta_i \tag{11a}$$
$$+ \sum_{i\in\mathcal{I}}\sum_{b\in\mathcal{B}}\sum_{c\in\mathcal{C}} \varpi^i_{b,c}\varepsilon^{R->C}(b,c)\theta_i$$
$$+ \sum_{i\in\mathcal{I}} \left(D^{\text{DU}}_i + D^{\text{CU}}_i + D^{\text{Core}}_i\right)$$
$$\text{s. t. } \frac{\alpha^a_i}{\rho^L_a \times F^L - \sum_{i\in\mathcal{I}} \alpha^a_i\theta_i} \le D^{\text{DU}}_i, \forall i \in \mathcal{I},\ a\in\mathcal{A}, \tag{11b}$$
$$\frac{\beta^b_i}{\rho^R_b \times F^R - \sum_{i\in\mathcal{I}} \beta^b_i\theta_i} \le D^{\text{CU}}_i, \forall i \in \mathcal{I},\ b\in\mathcal{B}, \tag{11c}$$
$$\frac{\gamma^c_i}{\rho^C_c \times F^C - \sum_{i\in\mathcal{I}} \gamma^c_i\theta_i} \le D^{\text{Core}}_i, \forall i \in \mathcal{I},\ c\in\mathcal{C}, \tag{11d}$$
$$(1), (2), (3), (4), (5).$$

We notice that the solution of (10) and (11) are identical since at the optimal point, the inequalities in (11b)-(11d) are active. In constraint in (11b), if α_i^a is zero, the constrain is always valid. If $\alpha_i^a = 1$, then the constraint becomes a linear constraint if we consider $1/D_i^{DU}$ as a new variable. First, we convert this constrain by adding a new variable M which is large and multiplied by $1 - \alpha_i^a$. For case of $\alpha_i^a = 0$ and $\alpha_i^a = 1$, the following constraint acts similar to (11b)

$$\frac{1}{D_i^{\text{DU}}} - (1 - \alpha_i^a)M \leq \rho_a^L \times F^L - \sum_{i \in \mathcal{I}} \alpha_i^a \theta_i, \tag{12}$$

where M is a large number. Similarly, we have the followings for (11c) and (11d)

$$\frac{1}{D_i^{\text{CU}}} - (1 - \beta_i^b)M \quad \rho_b^R \times F^R \times F_b^L - \textstyle\sum_{i \in \mathcal{I}} \beta_i^b \theta_i, \tag{13}$$

$$\frac{1}{D_i^{\text{Core}}} - (1 - \gamma_i^c)M \qquad \leq \qquad \rho_c^C \times F^C - \sum_{i \in \mathcal{I}} \gamma_i^c \theta_i, \tag{14}$$

Making a change of variables, as $\tilde{D}_i^{\text{DU}} = 1/D_i^{\text{DU}}$, $\tilde{D}_i^{\text{CU}} = 1/D_i^{\text{CU}}$, $\tilde{D}_i^{\text{Core}} = 1/D_i^{\text{Core}}$. Then, the problem can be written as

$$\begin{aligned} \min \quad & \sum_{i \in \mathcal{I}} \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} \varpi_{a,b}^i \varepsilon^{L->R}(a,b)\theta_i \\ & + \sum_{i \in \mathcal{I}} \sum_{b \in \mathcal{B}} \sum_{c \in \mathcal{C}} \varpi_{b,c}^i \varepsilon^{R->C}(b,c)\theta_i \\ & + \sum_{i \in \mathcal{I}} \frac{1}{\tilde{D}_i^{\text{DU}}} + \frac{1}{\tilde{D}_i^{\text{CU}}} + \frac{1}{\tilde{D}_i^{\text{Core}}} && (15) \\ \text{subject to} \quad & (1), (2), (3), (4), (5), (12), (13), (14). && (16) \end{aligned}$$

This is a problem with convex objective and linear binary constraints which can be solved using standard math tools.

5. Splitting Service Flow Traffics Along Servers

In this section, we formulate minimum delay CNF placement in O-RAN system but we allow the traffic of each SFR to split and traverse the RAN and core in multiple paths. This means, traffic flow of an SFR leaving an RRH may split into several paths where each path follows different routes between CU-CNF and DU-CNFs. Let $\pi^E_{a,i}$ denote the rate of SFR i processed at DU-CNF of edge sever $a \in \mathcal{A}$. Since θ_i is the total rate of SFR i, we have

$$\theta_i = \sum_{a \in \mathcal{A}} \pi^E_{a,i}, \tag{17}$$

Let $\pi^L_{b,i}$ and $\pi^C_{b,i}$ denote the rate of SFR i processed at CU-CNF of local sever $b \in \mathcal{B}$ and at Core CNF of core server $c \in \mathcal{C}$. For the local and core data centers, we similarly have

$$\begin{aligned} \theta_i &= \sum_{b \in \mathcal{B}} \pi^L_{b,i}, \\ \theta_i &= \sum_{c \in \mathcal{C}} \pi^C_{c,i}. \end{aligned} \tag{18}$$

We next determine the rate at which departing rate of edge data center a enters local data center b and then core data center c. Let $r^{E->L}(i,a,b)$ and $r^{L->C}(i,b,c)$ respectively denote the rate at which flow of SFR i moves from edge server a to local server b and moves from local server b to core server c. Flow conservation constraint mandates that

$$\begin{aligned} \pi^E_{a,i} &= \sum_{b \in \mathcal{B}} r^{E->L}(i,a,b), \\ \pi^L_{b,i} &= \sum_{a \in \mathcal{A}} r^{E->L}(i,a,b). \end{aligned} \tag{19}$$

And similarly

$$\begin{aligned} \pi^L_{b,i} &= \sum_{c \in \mathcal{C}} r^{L->C}(i,b,c), \\ \pi^C_{c,i} &= \sum_{b \in \mathcal{B}} r^{L->C}(i,b,c). \end{aligned} \tag{20}$$

The delay functions similarly can be written as

$$D^{\text{DU}}_{i,a} = \frac{1}{S^L_a \times F^L_a - \sum_{i \in \mathcal{I}} \pi^L_{a,i}}, \quad \text{if } \pi^E_{a,i} > 0. \tag{21}$$

$$D^{\text{CU}}_{i,b} = \frac{1}{S^R_b \times F^R_b - \sum_{i \in \mathcal{I}} \pi^R_{b,i}}, \quad \text{if } \pi^L_{b,i} > 0. \tag{22}$$

$$D^{\text{Core}}_{i,c} = \frac{1}{S^C_c \times F^C_c - \sum_{i \in \mathcal{I}} \pi^C_{c,i}}, \quad \text{if } \pi^C_{c,i} > 0. \tag{23}$$

Assuming constant transmission delay for the link between data centers. Let $\varepsilon^{L->R}(a,b)$ and $\varepsilon^{R->C}(b,c)$ denote the cost of using links from edge data center a to local data center b and from local data center b to core data center c, respectively. The expected delay for SFR i is the average delay on all paths which is obtained as

$$\begin{aligned} D_i \;&= \varepsilon^{L->R}(a,b) r^{L->R}(a,b) + \varepsilon^{R->C}(b,c) r^{R->C}(b,c) \\ &+ \sum_{a_{\mathcal{A}}} \frac{\pi^E_{a,i}}{S^L_a \times F^L_a - \sum_{i\in\mathcal{I}} \pi^E_{a,i}} \\ &+ \sum_{b_{\mathcal{B}}} \frac{\pi^R_{b,i}}{S^R_b \times F^R_b - \sum_{i\in\mathcal{I}} \pi^L_{b,i}} \\ &+ \sum_{c_{\mathcal{C}}} \frac{\pi^C_{c,i}}{S^C_c \times F^C_c - \sum_{i\in\mathcal{I}} \pi^C_{c,i}}. \end{aligned} \tag{24}$$

The minimum delay problem can be written as

$$\min_{\pi,r} \qquad D_{\text{total}} = \sum_{i\in\mathcal{I}} D_i \tag{25}$$

$$\text{subject to} \quad (17), (18), (19), (20). \tag{26}$$

$$\pi^E_{a,i}, \pi^L_{b,i}, \pi^C_{c,i} \geq 0, \forall i \in \mathcal{I}, a \in \mathcal{A}, b \in \mathcal{B}, c \in \mathcal{C}$$

$$r^{L->R}(i,a,b), r^{L->C}(i,b,c) \geq 0,$$

$$\forall i \in \mathcal{I}, a \in \mathcal{A}, b \in \mathcal{B}, c \in \mathcal{C}$$

6. Gradient Based Minimum Delay Algorithm

Let $\mathcal{P}_i$ denote the set of all paths of SFR i. This is the set of possible paths starting from an RRH, passing edge and local data centers and ends in a core data center. Let ϕ^p_i denote the rate of SFR i on path $p \in \mathcal{P}_i$. Let ψ_i denote the vector of rates ϕ^p_i as $\psi_i = [\phi^1_i, \ldots, \phi^{|\mathcal{P}_i|}_i]$ and ψ denote the rates of all SFRs on all paths. Let $\bar{\phi}^p_i$, $\bar{\psi}_i$, $\bar{\psi}$ denote the optimal values of ϕ^p_i, ψ_i, ψ in Problem (25), respectively. The following theorem describes the optimal point of this problem.

Theorem 1. *For the optimal point $\bar{\theta}^k_i$, we have*

$$\frac{\partial D(\psi)}{\partial \phi^p_i}\Big|_{\bar{\psi}} = \frac{\partial D(\psi)}{\partial \phi^q_i}\Big|_{\bar{\psi}}, \text{ if } \bar{\phi}^p_i \text{ and } \bar{\phi}^q_i > 0. \tag{27}$$

Proof. We use the proof by contradiction. Assume that these two terms are not equal and without loss of generality, assume that $\frac{\partial D(\psi)}{\partial \phi_i^p}|_{\bar{\psi}} > \frac{\partial D(\psi)}{\partial \phi_i^q}|_{\bar{\psi}}$. If we reduce small value of δ from $\bar{\phi}_i^p$ (assuming that $\bar{\phi}_i^p > 0$) and adds it to $\bar{\phi}_i^q$ while the rest of the data rates are constant, the solution remains feasible. Let $\tilde{\psi}_i$ denote the updated vector. The change vector for SFR i is $\delta_i = [\delta_i^1, \dots, \delta_i^{|\mathcal{P}_i|}]$ where all δ_i are zero except $\delta_i^q = -\delta_i^p = \delta$. The total change vector is $\Delta = [\delta_1, \dots, \delta_{|\mathcal{I}|}]$. For small values of δ, the change in D can be first order approximated as

$$\begin{aligned} D(\tilde{\psi}) - D(\bar{\psi}) &= \sum_{i \in \mathcal{I}} D_i(\tilde{\psi}_i) - D_i(\bar{\psi}_i) \\ &\approx \nabla D \times \Delta = \frac{\partial D(\psi)}{\partial \phi_i^p}|_{\bar{\psi}} \delta_i^p + \frac{\partial D(\psi)}{\partial \phi_i^q}|_{\bar{\psi}} \delta_i^q \\ &= \delta \left(-\frac{\partial D(\psi)}{\partial \phi_i^p}|_{\bar{\psi}} + \frac{\partial D(\psi)}{\partial \phi_i^q}|_{\bar{\psi}} \right). \end{aligned} \tag{28}$$

Since we assume that $\frac{\partial D(\psi)}{\partial \phi_i^p}|_{\bar{\psi}} > \frac{\partial D(\psi)}{\partial \phi_i^q}|_{\bar{\psi}}$, the last term in (28) is negative which means $D(\tilde{\psi}) - D(\bar{\psi}) < 0$. It contradicts with optimality of $\bar{\psi}$. If we repeat the case for $\frac{\partial D_i}{\partial \phi_i^p}|_{\bar{\psi}} < \frac{\partial D_i}{\partial \phi_i^q}|_{\bar{\psi}}$, we obtain the same results. Therefore, at the optimal point, (27) is valid. ■ Theorem 1 suggests an iterative gradient based algorithm as listed in Algorithm 1. The algorithm iterates on all the SFRs. For each SFR, we deduct rate from a path with the highest derivative and we add it to the path with the lowest derivative. It makes sure that it is a descent direction which results in lower value of the objective function.

7. Numerical Results

In this section, we evaluate the performance of the proposed solutions. We compare these schemes with varying the capacity of data centers and varying the number of SFRs.

We consider a base capacity for the data centers of each region. Table 2 shows the amount of *base* resources in each DC. First, we vary the number of SFRs for GBMD(Gradient-Based Minimum Delay) method while the capacity of data centers are fixed. Fig. 3 shows the average aggregated end-to-end delay of

Algorithm 1: Gradient Based Minimum Delay (GBMD) Algorithm

```
Start with a feasible solution for $\phi_i^p, \forall i \in \mathcal{I}, p \in \mathcal{P}_i$. **while** *Not Converged* **do**
    **for** *any SFR $i \in \mathcal{I}$* **do**
        Find path $p$ of SFR $i$ with highest $\frac{\partial D}{\phi_i^p}$,
        Find path $q$ of SFR $i$ with minimum $\frac{\partial D}{\phi_i^q}$,
        if $\phi_i^p > 0$,
        Set: $\phi_i^p = \phi_i^p - \delta$.
        Set: $\phi_i^q = \phi_i^q + \delta$.
        Check if new values of $\phi_i^p$ and $\phi_i^q$ are feasible. If not, reverse last steps.
    **end**
**end**
```

GBMD algorithm by varying the number of SFRs. The figure contains three curves for base, 1.2 times the base, and 1.5 times the base capacity of DCs as shown in Table II. It is shown that by increasing the number of SFRs, the end-to-end delay increases. Increasing the amount of resources reduces the minimum delay as we expect. It is important to notice that in GBMD method, SFR rate will be splitted between possible Local DCs; here, total rate of each SFR is considered as 300Mbps. Figures 5 and 4 compare the two proposed schemes in terms of the average aggregated end-to-end delay. *Single path* refers to the problem modeled in (15). We compare it with GBMD algorithm in Algorithm 1. GBMD algorithm achieves lower delay since it has a wider feasible region compared to the single path algorithm. For a fixed amount of data center resources, the delay of single path increases exponentially while it is more close to linear for GBMD.

For the next simulation, we vary the data center capacity. As it is mentioned in Table II, we consider different base values for Local, Regional, and Core data Centers. For this simulation, we increase the total amount of capacity for all

Table 2: Data Centers Resource Settings

DC type	Processing capacity (no. of CPU cores)	Transmission rate (Gbps)	Storage (TB)
local	1,000	1,000	2,000
Regional	5,000	1,000	3,000
Core	8,000	5,000	4,000

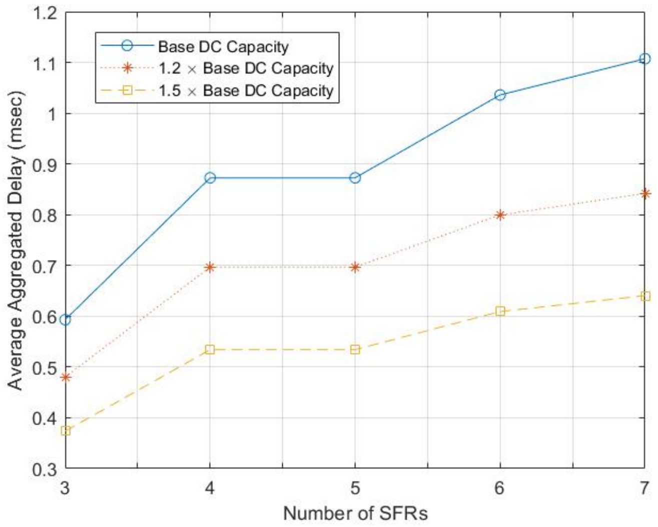


Figure 3: The aggregated end-to-end delay versus number of SFRs for base, 1.2 times base, and 1.5 times base capacity of DCs.

layers. More precisely, at each step, we add 100GB to the processing capacity of each data center at all layers. In this simulation, there are four local, four regional and one core data centers. Therefore, in each step, we add 900GBs to the total capacity. The results are shown in Fig. 5 . Increasing the capacity of the data centers reduces the aggregate end-to-end delay substantially. This suggests that by marginally increasing the capacity, one exponentially gains in lower end-to-end delay. Since CAPEX and OPEX is associated by increasing

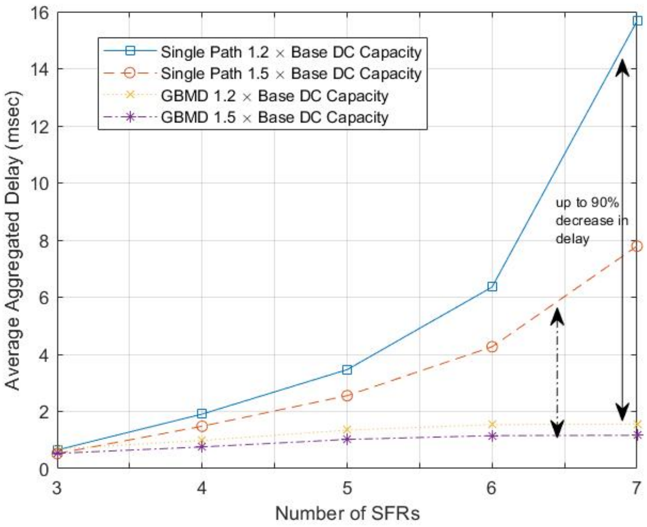


Figure 4: The aggregated end-to-end delay versus variable number of SFRs for different amount of data center resources.

the amount of resources, there is inherently a tradeoff between delay and cost in choosing the right amount of data center resources.GBMD algorithm has a direct impact on resource allocation by decreasing E2E delay up to 72 percent.

While supporting different QCIs (Quality Class Identifier), which is a mechanism used in 3GPP Long Term Evolution (LTE) networks to ensure bearer traffic is allocated appropriate Quality of Service (QoS), is important for network providers, we have imposed QoS parameters on some service inputs(in terms of GBR(Guaranteed Bit Rate)) and a comparison between best effort and QoS based conditions shows that adding QoS will increase e2e Delay in case of limited resources. This is a trade-off between providing QoS based services for high value customers and providing best effort services for all levering GBMD algorithm. The simulation result is shown if Fig.6.

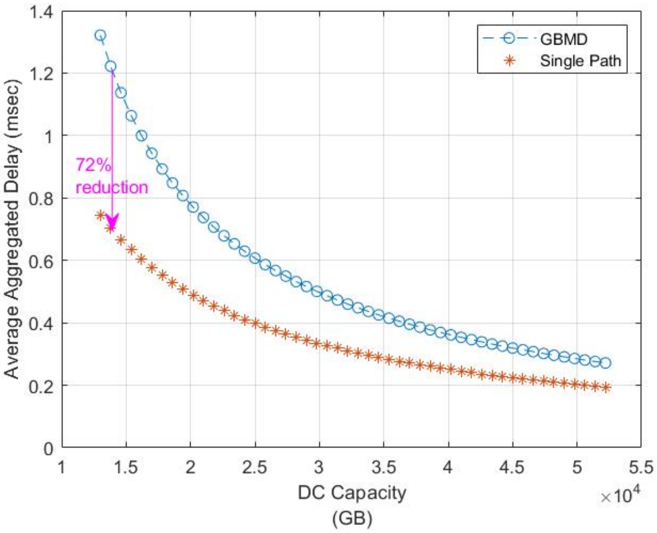


Figure 5: The aggregated end-to-end delay versus variable amount of data center resources.

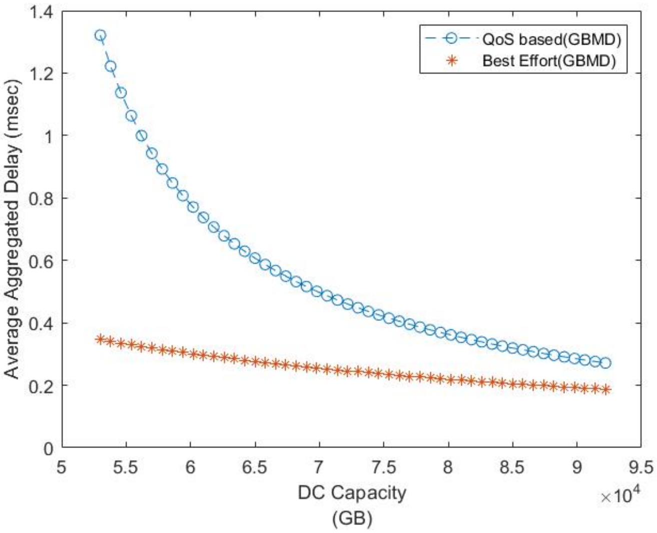


Figure 6: The Comparision between QoS-based implementation and Best effort approach levering GBMD algorithm in a network.

8. Conclusion

Architectural transformation is an inevitable step towards meeting 5G use case requirements. In this paper, end to end virtualization in access, IPBB, and core network using local, regional, and core data centers is studied. It is a solution to reduce the end-to-end delay of service requests by offering flexibility and increasing chance of selecting proper resources. O-RAN proposes the use of open and virtualized RAN. In O-RAN, different layers of 5G RAN can run on different data centers at different locations as VNFs/CNFs. A chain of those VNFs/CNFs provides the basic telecom connectivity for a 5G network. A hierarchical three layer data center architecture is deployed. To support multi-tenancy, a new chain of CNFs is created for each SFR. It provides isolation between the service management. We first model the CNF placement and resource allocation problem for the case that the traffic of an SFR traverses one chain via a single traffic path. Since it is not tractable, we model the case that the traffic of an SFR can traverse multiple chains. We propose a gradient based method which can efficiently find the optimal solution. Using numerical results, we investigate the performance of the proposed algorithm for varying number of SFRs and varying capacity of data center resources. We show that increasing the data center capacity exponentially reduces the end-to-end delay.

References

[1] W. Kiess, X. An, S. Beker , ”Software-as-a-Service for the Virtualization of Mobile Network Gateways” *GLOBECOM-IEEE Global Communications Conference,* San Diego, CA, USA, 6-10 Dec. 2015.

[2] L. J. Vora, ”Evolution of mobile generation technology: 1G to 5G and review of upcoming wireless technology 5G,” *IJMTER-International Journal of Modern Trends in Engineering and Research,*, Vol.2, no. 10, pp. 281-290, Oct. 2015.

[3] E. Coronado, S. N. Khan, and R. Riggio, ”5G-empower: A software-defined networking platform for 5G radio access networks,” *IEEE Transactions on Network and Service Management,* vol. 16, no. 2, pp. 715728, June, 2019.

[4] Md. Faizul Bari, ”Resource Orchestration in Softwarized Networks”. *University of Waterloo*, PhD Thesis, *Waterloo,* Ontario, Canada, 2018.

[5] W. Xia, T. Q. S. Quek, J. Zhang, S. Jin, and H. Zhu, ”Programmable hierarchical C-RAN: From task scheduling to resource allocation,” *IEEE Transactions on Wireless Communications,* vol. 18, no. 3, pp. 2003-2016, March, 2019.

[6] V. G. Nguyen, A. Brunstrom, K. J. Grinnemo, and J. Taheri, ”SDN/NFV-Based Mobile Packet Core Network Architectures: A Survey,” *IEEE Communication Surveys and Tutorials,* vol. 19, no. 3, April, 2017.

[7] S. Li, L. D. Xu, S. Zhao, ”5G Internet of Things: A Survey,” *Georg-August-Universitt Gttingen*, Journal of Industrial Information Integration, Elsevier, Vol. 10, pp. 1-9, June, 2018.

[8] Q. Duan, N. Ansari, and M. Toy, ”Software-Defined Network Virtualization - An Architectural Framework for Integrating SDN and NFV for Service Provisioning in Future Networks,” *IEEE Network,* vol. 30, no. 5, pp. 1016, Sept. 2016.

[9] ETSI, ”ETSI GS NFV-IFA 010 V3.3.1”. *European Telecommunications Standards Institute*, Sept. 2019.

[10] Open Source Cloud Computing Software,” http://openstack.org/ , July, 2016.

[11] S. G. Kulkarni, ”Resource Management for Efficient, Scalable and Resilient Network Function Chains,” , PhD Thesis, Georg-August-Universitt Gttingen, 2018.

[12] , C-RAN: The road towards green RAN, *Mobile China, Hong Kong, China*, White Paper ver. 2, Oct. 2011.

[13] A. M. Mahmood , "A New Processing Approach for Reducing Computational Complexity in Cloud-RAN Mobile Networks" *IEEE Access*, vol. 6, pp. 6927 - 6946, Dec. 2017.

[14] F. Tonini, C. Raffaelli, L. Wosinska, and P. Monti, "Cost-optimal deployment of a C-RAN with hybrid fiber/FSO fronthaul," *Journal of Optical Communications and Networking,* vol. 11, no. 7, pp. 397408, June. 2019.

[15] D. Harutyunyan and R. Riggio , "How to migrate from operational lte/lte-a networks to C-RAN with minimal investment?" *IEEE Transactions on Network and Service Management,* vol. 15, no. 4, pp. 15031515, Dec. 2018.

[16] J. Tang, W. P. Tay, and T. Quek, "Cross-layer resource allocation with elastic service scaling in cloud radio access network" *IEEE Transactions on Wireless Communications,* vol. 14, no. 9, pp. 50685081, Sept. 2015.

[17] K. Wang, K. Yang and C.S. Magurawalage, "Joint energy minimization and resource allocation in C-RAN with mobile cloud," *IEEE Transactions on Cloud Computing*, vol. 6, no. 3, pp. 760770, Sept. 2018.

[18] ORAN Alliance, "O-RAN: Towards an Open and Smart RAN". *White paper*, 2018.

[19] J. P. Garzon, "Architecture, Modeling, Planning, and Dynamic Provisioning of Softwarized 5G mobile Core Networks," PhD Thesis, *Universitas Granatensis*, 2018.

[20] M. Yang, Y. Li, D. Jin, L. Su, S. Ma, and L. Zeng, "OpenRAN: A Software-defined RAN Architecture Via Virtualization," *in Proc. of ACM SIGCOMM,* Hong Kong, China, 12-16 Aug. 2013.

[21] Y. D. Beyene, R. Jantti, and K. Ruttik, "Cloud-RAN Architecture for Indoor DAS," *IEEE Access* , vol. 2, pp. 1205-1212, Oct. 2014.

[22] RAJAGOPAL, "System and Method for reduction in fronthaul interface bandwidth for CloudRAN," *United States, Patent Application Publication, US 2019 / 0289497 A1*, March, 2019.

[23] L. Gavrilovska, V. Rakovic,and D. Denkovski, "From Cloud RAN to Open RAN," *Springer, Wireless Pers Communication,* pp.15231539, March, 2020.

[24] T. Koponen *et al.*, "Network Virtualization in Multi-tenant Datacenters," *Proceedings of the 11th USENIX Conference on Networked Systems Design and Implementation, USENIX Association,*Berkeley, CA United States , Berkeley, CA, United States, pp.203216, April, 2014.

[25] K.Katsalis, N. Nikaein, E. J. Schiller, A. Ksentini, and T. Braun, "Network Slices toward 5G Communications: Slicing the LTE Network," *IEEE COMMUN MAG-IEEE Communications Magazine,* vol. 55, no.8, pp.146 - 154, Aug., 2017.

[26] V. Tikhvinskiy and G. Bochechka, "Prospects and QoS Requirements in 5G Networks," *Journal of Telecommunications and Information Technology*, pp.23-26, Jan., 2015.

[27] Z. Tan,C. Yang, J. Song, Y. Liu, and Z. Wang, "Energy consumption analysis of C-RAN architecture based on 10g epon front-haul with daily user behaviour", *ICOCN-14th IEEE International Conference on Optical Communications and Networks*, Nanjing, China, pp. 1-3, July, 2015.