

Energy-Efficient Virtual Base Station Formation in Optical-Access-Enabled Cloud-RAN

Xinbo Wang, Saigopal Thota, Massimo Tornatore, Hwan-Seok Chung, Han-Hyub Lee, Soomyung Park, Biswanath Mukherjee, *Fellow, IEEE*

Abstract—In recent years, the increasing traffic demand in radio access networks (RAN) has led to considerable growth in the number of base stations (BS), posing a serious scalability issue, including the energy consumption of BSs. Optical-access-enabled Cloud-RAN (CRAN) has been recently proposed as a next-generation access network. In CRAN, the digital unit (DU) of a conventional cell site is separated from the radio unit (RU) and moved to the “cloud” (DU cloud) for centralized signal processing and management. Each DU/RU pair exchanges bandwidth-intensive digitized baseband signals through an optical access network (fronthaul). Time-Wavelength Division Multiplexing (TWDM) Passive Optical Network (PON) is a promising fronthaul solution due to its low energy consumption and high capacity. In this study, we propose and leverage the concept of a Virtual Base Station (VBS), which is dynamically formed for each cell by assigning virtualized network resources, i.e., a virtualized fronthaul link connecting the DU and RU, and virtualized functional entities performing baseband processing in DU cloud. We formulate and solve the VBS formation (VF) optimization problem using an Integer Linear Program (ILP). We propose novel energy-saving schemes exploiting VF for both the network planning stage and traffic engineering stage. Extensive simulations show that CRAN with our proposed VF schemes achieves significant energy savings compared to traditional RAN and CRAN without VF.

Index Terms—Optical access, TWDM-PON, CRAN, fronthaul, DU cloud, Energy efficiency.

I. INTRODUCTION

TRAFFIC demand in radio access networks (RAN) is increasing rapidly, forcing mobile operators to provide larger capacity to serve more users. So the number of base stations (BSs) is expected to grow dramatically, posing a serious scalability issue, including the energy efficiency of BSs, which are responsible for more than half of the energy consumed in RANs [1]. For example, from 2004 to 2009, the number of BSs deployed by China Mobile grew from 200,000 to 500,000, resulting in a two-fold increase of power consumption, 70% of which came from BSs [2].

In traditional distributed RAN (DRAN), each BS consists of two components, co-located in the same cell site, the Digital Unit (DU) or Baseband Unit (BBU), and the Radio Unit (RU) or Remote Radio Head (RRH). In the BS, the DU is responsible for baseband processing, while the RU is responsible for

transmitting/receiving and digitizing radio signals. However, traffic load in the RAN also varies significantly during a day, and the capacity of DU is designed for peak load. Within the same BS, the baseband processing resources of a DU are dedicated to the associated RU, so the unused resources of the DU cannot be shared by RUs in other BSs; thus, all DUs in the RAN need to remain active, causing energy waste in RAN during off-peak periods. Moreover, each BS needs independent energy-consuming “housing” facilities (cooling systems, etc.) for its DU, and all these energy-consuming components need to stay active all the time. Therefore, DRAN does not represent a future-proof and scalable solution for next-generation RAN.

To address these scalability issues, a new architecture, called Cloud-RAN (CRAN) [2]–[4], has been recently proposed. CRAN centralizes DUs in a single location and leaves RUs at cell sites. CRAN can achieve cost savings by allowing smaller footprint, less power consumption of outdoor equipment, and shared infrastructure in the “cloud”. But traditional CRAN employs DU architectures, such as DU hoteling/pooling, where the DUs are co-located, but remain separate and are each individually connected to RUs. Currently, the DU architecture that indeed fulfills the “cloud” RAN is the DU cloud, where DUs are implemented on general-purpose servers that can be flexibly configured [5]. By virtualizing the computing resources of each DU, DU can be dynamically shared by multiple cells. CloudIQ [6] showed that pooling resources can reduce the overall cost of a network, however it relied on statically allocating resources (rather than dynamically reallocating resources according to varying loads), and it did not consider energy saving in CRAN. Ref. [7] proposed DU virtualization and modeled it as a bin-packing problem. Ref. [8] proposed a virtualization approach for DU cloud, where baseband processing in a DU is virtualized as functional entities, cell processing (PHYcell) and user processing (UP), both of which can be virtualized and reconfigured in a DU for a cell and a user, respectively. This concept brings new opportunities for energy saving in CRAN.

To connect the DU cloud and RUs, CRAN requires a high-capacity access network, called “fronthaul” [5]. As opposed to the backhaul in traditional RAN, which carries Layer-3 packets to the core Internet, the fronthaul transports digitized baseband signals. Optical fronthaul solutions can include dedicated fibers, wavelength-division multiplexing (WDM) passive optical network (PON), time-wavelength division multiplexing (TWDM) PON [10]. High cost of the first solution might jeopardize the cost savings offered by CRAN. Therefore, TWDM-PON has been shown to be a promising fronthaul

X. Wang, S. Thota, M. Tornatore, and B. Mukherjee are with the Department of Computer Science, University of California Davis, Davis, CA, 95616 USA e-mail: xbwang@ucdavis.edu.

M. Tornatore is also with the Department of Electronics, Information and Bioengineering Politecnico di Milano, Italy.

H. Chung, H. Lee, and S. Park are with ETRI Korea.

Manuscript received March 15, 2015; revised May 15, 2015 and September 1, 2015.

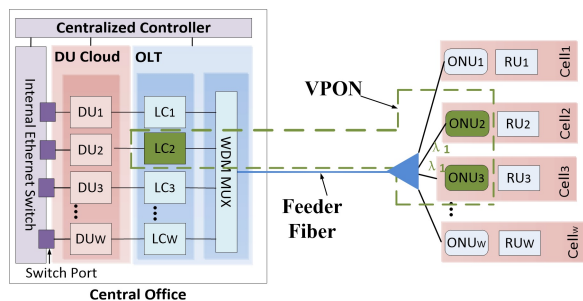


Fig. 1. Illustration of TWDM-PON-enabled CRAN architecture.

solution for CRAN, thanks to its low cost, low energy consumption, and abundant bandwidth. Here, we consider that TWDM-PON is a plausible technology as (almost) fixed bandwidth allocation is assumed and its latency meets the requirements of fronthauling (see, e.g., [10]). Refs. [11], [12] describe approaches to save energy in TWDM-PON. In [13], enabling technologies are proposed to save energy in all-optical networks. However, these energy-saving approaches cannot directly apply to CRAN. In [14], an energy-efficient optimization problem with resource assignment and power allocation in CRAN is proposed. In [15], an energy-efficient DU cluster testbed for CRAN is developed, but it did not consider the energy saving in fronthaul.

We summarize our contributions in this work as follows. We propose CRAN architecture consisting of a TWDM-PON-enabled fronthaul, DU cloud, an internal Ethernet switch that can redirect traffic in DU cloud, and a centralized controller that can schedule network resources, as shown in Fig. 1. We consider that network resources in DU cloud and TWDM-PON can be reconfigured and virtualized. In particular, we abstract base stations (BSs) in CRAN as multiple virtual base stations (VBSs) comprised of virtualized network resources (e.g., a virtualized fronthaul link and virtualized baseband processing). The centralized controller takes all decisions about how to form VBS and which cell is served by which VBS according to varying traffic loads in the CRAN. By efficiently forming VBS for each cell, active network resources can be minimized by shutting down unnecessary physical resources. Our previous work studied energy saving via VBS formation (VF) under static traffic provisioning [16]. In this work, we present a comprehensive study on using VF to save energy in CRAN. We formulate and solve the VF optimization problem using an Integer Linear Program (ILP). A static VF scheme is proposed for the network planning stage. For the traffic engineering stage, we propose dynamic VF schemes that can take online decisions for VF. Extensive simulations show that significant energy savings can be achieved by CRAN with VF compared to DRAN and CRAN without VF. Performances of dynamic VF schemes are also evaluated and analyzed.

The rest of the study is organized as follows. Section II introduces the CRAN system architecture that enables VF. Section III proposes the energy-saving model. In Section IV, we formulate the VF optimization problem using an ILP. We design heuristic algorithms for the static and the dynamic cases in Section V. Numerical evaluation is given in Section VI, and

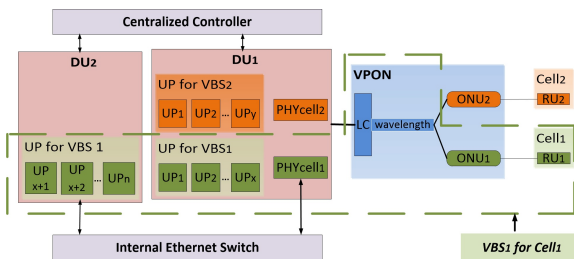


Fig. 2. Illustration of a Virtual Base Station (VBS).

Section VII concludes this work.

II. SYSTEM ARCHITECTURE

A. Introduction to CRAN

Recently, the new CRAN architecture is gaining attention because of its great potential. As shown in Fig. 1, the digital units (DUs), which are responsible for electronic baseband processing, can be moved from the distributed BSs to a DU cloud at the central office, serving a large group of radio units (RUs), which are simplified as radio-frequency (RF) electronics. The DU cloud and RUs are connected by high-speed digital fronthaul links transmitting digitized signals using the Common Public Radio Interface (CPRI) [17].

In [18], the authors provide a detailed description for fronthaul requirement in CRAN. First, macro cells typically yield up to 15 RUs per site. Second, CPRI is a constant bit-rate interface, which requires data rates from 614.4 Mbps up to 10.137 Gbps with bit error ratio on the fronthaul link lower than 10^{-12} . Third, the baseband processing at DU should be less than 1 ms. The round-trip time dedicated to fronthaul and to the optical network segment should be less than $500 \mu\text{s}$. Finally, the CPRI link should contribute less than ± 2 ppb (parts per billion) to the overall frequency accuracy requirement on the air interface for LTE. A detailed analysis on feasibility of the proposed architecture is an open problem, but the requirements in terms of latency, capacity, etc. are expected to be compatible with our proposed architecture.

Although it has high requirements on fronthaul, CRAN has benefits that can justify its realization costs. First, CRAN is a cost-saving architecture compared to DRAN as it simplifies network operation, maintenance, and upgrade by concentrating in a single location for RAN processing and coordination equipment. Specifically, CRAN has lower cost at cell sites, such as smaller footprint and power consumption of outdoor equipment, and can dynamically share infrastructure in DU cloud, according to the traffic variation. In addition to cost-related benefits, CRAN also eases implementation of advanced radio-coordination techniques, such as Coordinated Multipoint (CoMP), to improve RAN coverage, bitrate, and throughput. At present, CRAN can be readily supported by fronthaul links that transmit CPRI signals over either dedicated fibers or over time-wavelength multiplexing division (TWDM) passive optical networks (PON) for distances up to 40 km and various loss budgets [5], [19]. The dedicated-fiber fronthaul solution is too costly and may nullify the cost-saving gains of CRAN.

We consider TWDM-PON as the most cost-effective solution that is viable in today's networks to meet these requirements¹.

B. TWDM-PON-Enabled CRAN Architecture

We propose the CRAN architecture using TWDM-PON as the fronthaul (as in [10]), and using the DU cloud virtualization solution (as in [8]), which is implemented on general-purpose servers that can be dynamically reconfigured. As depicted in Fig. 1, an RU receives RF signals from all users in the cell. For every cell, there is one Optical Network Unit (ONU) connected to the RU serving the cell (for simplicity, we assume one RU per cell). The ONU uses CPRI to send packets containing the digitized baseband signals to the DU cloud, and thanks to the high capacity of a wavelength channel, multiple ONUs can share the same wavelength in the TWDM-PON. At the Optical Line Terminal (OLT) of PON, every wavelength is served by a line card (LC) (i.e., transceiver). The combination of an LC and multiple ONUs communicating over the same wavelength is viewed as an independent PON, called Virtual PON (VPON), in which multiple ONUs share a wavelength in time-division multiplexing (TDM) manner. So, a TWDM-PON can provide multiple VPONs over different wavelengths.

The DU cloud and OLT are co-located at the Central Office (CO). In the DU cloud, each DU is associated with one LC, and provides baseband processing functions for cells associated with the VPON. Thus, the set of DUs and LCs have the same cardinality (W). Due to the flexibility of TWDM-PON (to provide dynamic wavelength assignment and sharing), the association of a RU to its corresponding DU can be reconfigured, without the need of putting any active switch between the DUs and the LCs. For example, in Fig. 1, RU_2 and RU_3 are associated with DU_2 by configuring their serving wavelength as λ_1 . An internal Ethernet switch is put behind the DU cloud. There should be W ports on the switch, each connecting to a DU. A port will be activated when the associated DU needs to redirect traffic to/from other DU. As the source and destination of this redirection are fixed for a time duration, we can use the cut-through switch [21] to quickly forward the packet without storing the whole packet. The latency of a cut-through switch is 1-20 μ s, which is small compared to latency requirement of baseband processing (1 ms).

C. Virtual Base Station

Network resources in the proposed architecture, including baseband processing, switch ports, and LCs, can be virtualized respectively in DU cloud, switch, and TWDM-PON. A centralized controller flexibly configures these virtualized resources by sending instructions to their physical hardware.

For baseband processing in a DU, it is virtualized as two types of functional entities: cell processing (PHYcell) and

user processing (UP) [8]. A DU can set up a PHYcell for each RU (cell) served by it (see Fig. 2). The PHYcell process cell control messages and multiplex/de-multiplex user traffic belonging to a RU. A cell can change its serving DU by *PHYcell shifting*, which tears down the current PHYcell at a DU and sets up a new PHYcell in another DU. PHYcell shifting can be achieved by changing the serving VPON (i.e., retuning the serving wavelength). Also, a DU will set up a UP for every user served by the RU that has PHYcell. UP provides baseband processing for user traffic. A UP can be redirected in DU cloud through *UP redirection*. When a DU is overloaded, it will redirect some UPs to other DUs with extra baseband processing resources through the internal switch. For example, in Fig. 2, not all UPs of $PHYcell_1$ can be accommodated by DU_1 , so some of them are redirected to DU_2 .

For port resources in the switch, we assume that ports can be flexibly turned on/off for energy-saving purpose to connect source and destination DUs of UP redirection.

One TWDM-PON can support multiple VPONs, each of which is formed for a DU and can be shared by several RUs. By configuring the virtualized resources, we can form a virtual base station (VBS) for each cell, which consists of one PHYcell, multiple UPs (possibly distributed among one or more DUs), and a virtualized fronthaul link (connecting the RU and DU) provided over a VPON. As enclosed by the green dotted line in Fig. 2, VBS_1 is formed for $Cell_1$. By allocating "just enough" virtualized resources to each VBS, we can effectively utilize the shared network hardware, i.e., minimize the number of active DUs, switch ports, and LCs. By dynamic VBS formation (VF), we can exploit the temporal variation of traffic load in RAN to save energy in CRAN.

D. Other Architectures

To quantify the energy savings that can be achieved in this CRAN architecture by our proposed VF scheme, we compare it with two other architectures: DRAN and CRAN without VF. In DRAN, every DU is collocated with its RU at cell site. Hence, every cell needs its own housing facility, and DU should remain active all the time. In traditional CRAN, although DUs are co-located in the DU cloud, there is no sharing of DUs and LCs in DU cloud. So, every cell needs an active DU and an LC dedicated to serve it.

III. ENERGY CONSUMPTION MODELS

For CRAN, the energy-consumption terms come from the consolidated "housing" facilities, network devices in the CO, and network devices at cell sites. Housing facilities require an amount of power $P_{H,pool}$ to ensure proper operational conditions (e.g., cooling), although they do not perform network functions. Active network devices in CO include the internal switch, DUs, and LCs. Energy consumption of the switch is modeled as the sum of a baseline (traffic-independent) energy $P_{S,B}$ and a traffic-dependent energy $P_{S,P}$, increasing with number of active ports n_P (each active port is assigned to a DU which participates in UP redirection). For DU power consumption, we do not consider processor dynamic voltage and frequency scaling (DVFS) [22], so DU will consume a

¹There are various on-going projects aimed at providing low-latency PON for fronthaul, e.g., [10] proposed a mobile-DBA scheme based on the collaboration between the mobile network and the fronthaul network. The proposed scheme can reduce the latency to 50 μ s, and jitter to several tens of μ s. Work is in progress to study the feasibility of fast packet-switched PON for CPRI [20]. We do not restrict our architecture to one solution.

TABLE I
SUMMARY OF ENERGY MODEL PARAMETERS AND THEIR VALUES

Description	Parameter	Value
Switching: baseline consumption	$P_{S,B}$	50 W
Switching: per-port consumption	$P_{S,P}$	15 W/port
Housing: stand-alone DU	$P_{H,D}$	600 W
Housing: baseline of DU pool	$P_{H,pool}$	500 W
Housing: load-dependent of DU pool	P_{DU}	100 W/DU
Line card: per-line-card consumption	P_{LC}	5 W/LC
Per-RU consumption	P_{RU}	20 W/RRH
Per-ONU consumption	P_{ONU}	7.7 W/ONU

constant energy P_{DU} when it is active; otherwise, it consumes no energy. Similarly, an active LC consumes a constant energy P_{LC} . Active network devices at cell sites include ONUs and RUs. We do not consider the sleep mode of RU and ONU, which will not make much difference between CRAN and DRAN, and is not the focus in our study. Therefore, they need to remain active and consume a constant energy, P_{RU} and P_{ONU} , respectively, in all architectures. So, energy consumption for CRAN with VF architecture is:

$$P_{C1} = P_{H,pool} + P_{S,B} + P_{S,P} \cdot n_p + (P_{RU} + P_{ONU}) \cdot N + \sum_{m=1}^W (P_{LC} \cdot y_m + P_{DU} \cdot s_m) \quad (1)$$

where N is the number of cell sites, W is the number of LCs or DUs, and y_m and s_m are boolean variables indicating whether the m^{th} LC and DU are active, respectively (note that we can have a DU on even if the LC is off, but not vice versa).

For traditional CRAN, there is no cost of switch, but the energy consumption of all LCs and DUs must be always counted (i.e., it does not depend on traffic) as it does not have a sharing mechanism. Also, it consumes a constant energy from ONUs and RUs at cell sites. The model is as follows:

$$P_{C2} = P_{H,pool} + (P_{RU} + P_{ONU}) \cdot N + (P_{LC} + P_{DU}) \cdot W \quad (2)$$

For DRAN, a cell with DU consumes a constant energy, $P_{H,D}$, including energy consumed by DU and housing facilities, but there is no energy consumption from TWDM-PON system and internal switch at CO. So, the model is as follows:

$$P_D = (P_{ONU} + P_{H,D}) \cdot N \quad (3)$$

A summary of energy consumption values [11], [19], [23], [24] is given in Table I.

IV. VBS FORMATION OPTIMIZATION PROBLEM

A. Problem Statement

Let us first consider a simplified version of the problem where the bandwidth of users' requests is given and remains unchanged, and the bandwidth bottleneck is at the DU (it could also be at the wavelength)². In this static version, the

²We assume that the bandwidth by a DU from a computational point of view is less than the capacity of a wavelength. This is reasonable as a wavelength will always be sufficient for accommodating traffic in a DU.

VF problem shares similarities with the one-dimensional bin-packing (1D-BP) problem where the size of the bin is the bandwidth limit of each DU and the sizes of the items are the users' bandwidth demands, namely the bandwidth demand of UPs. Our goal is to use the minimum number of bins (DUs) to pack all the items (accommodate the demand of PHYcells and UPs). However, there are differences between bin-packing and VF problem. First, we have two types of items, UP and PHYcell, and cost of assigning UPs depends on assignment of the corresponding PHYcells because UP redirection needs to consume extra energy. Second, DU's (bin's) capacity can be categorized in two independent "sub-bins": DU's UP capacity in terms of supported users' bandwidth, and DU's PHYcell capacity in terms of number of cells. Finally, our goal is to minimize the overall energy consumption, instead of merely minimizing number of bins used as in Eqn. (1). So, assignment of items to bins involves two separate (but correlated) assignments: PHYcells must be assigned to DUs within the DU's PHYcell capacity constraint; UPs must be assigned preferably to the DU that hosts the PHYcell of the DU, but some DUs might be assigned to other DUs in case of overload. But we want to minimize the redirection because this will cause energy consumption. So, traditional approximation algorithms for 1D-BP cannot be used to solve our VF problem, and we need other methods for this static VF (S-VF) problem. Besides, when users dynamically arrive and hold service for a certain time, S-VF becomes an even harder dynamic VF (D-VF) problem. Thus, an efficient online algorithm is necessary for it. Below, we first formulate S-VF using an ILP; then we discuss S-VF and D-VF problems in Section V.

B. Modeling the VF Optimization Problem

Given: Network topology, users' traffic demands for all cells, wavelength capacity, DU's UP capacity in terms of bandwidth, DU's PHYcell capacity in terms of cell number, switch bandwidth, and energy consumption values.

Output: How many LCs, DUs and switch ports VBSs to use.

Objective: Minimize overall energy given by Eqn. (1).

1) Input parameters:

- W : set of wavelengths, which also indicates the set of LCs and the set of DUs (by using the same index, w).
- I : set of cells.
- J_i : set of users in cell i .
- C_W : capacity of a wavelength.
- C_{UP} : DU's UP capacity in terms of supported bandwidth (which is mapped from the processing capability).
- C_{PH} : DU's PHYcell capacity in terms of cell number.
- C_{ES} : capacity of the switch.
- $b_{i,j}$: traffic load of j^{th} user in i^{th} cell.
- B : a very big positive value.

2) Binary Variables:

- $x_{i,w}$: if cell i is served by wavelength w .
- y_w : if LC w is active.
- $u_{i,j,w}$: if UP of user j in cell i is served by DU w .
- $k_{i,j}$: if UP of user j in cell i has been redirected.
- r_w : if DU w participate in UP redirection.
- s_w : if DU w is active.

- e : if switch is active.
- $g_{i,j,w}$: if user j of cell i is a redirected UP in DU w .

3) Constraints:

$$\sum_{i \in I} x_{i,w} \leq C_{PH}, \forall w \in W \quad (4)$$

$$\sum_{i \in I} \sum_{j \in J_i} u_{i,j,w} \cdot b_{i,j} \leq C_{UP}, \forall w \in W \quad (5)$$

$$\sum_{i \in I} \sum_{j \in J_i} k_{i,j} \cdot b_{i,j} \leq C_{ES} \quad (6)$$

$$\sum_{w \in W} x_{i,w} = 1, \forall i \in I \quad (7)$$

$$\sum_{w \in W} u_{i,j,w} = 1, \forall i \in I, \forall j \in J_i \quad (8)$$

$$B \cdot y_w \geq \sum_{i \in I} x_{i,w}, \forall w \in W \quad (9)$$

$$y_w \leq \sum_{i \in I} x_{i,w}, \forall w \in W \quad (10)$$

$$B \cdot s_w \geq \sum_{i \in I} \sum_{j \in J_i} u_{i,j,w}, \forall w \in W \quad (11)$$

$$s_w \leq \sum_{i \in I} \sum_{j \in J_i} u_{i,j,w}, \forall w \in W \quad (12)$$

$$y_w \leq s_w, \forall w \in W \quad (13)$$

$$g_{i,j,w} \leq u_{i,j,w} + x_{i,w}, \forall w \in W, \forall i \in I, \forall j \in J_i \quad (14)$$

$$g_{i,j,w} \geq u_{i,j,w} - x_{i,w}, \forall w \in W, \forall i \in I, \forall j \in J_i \quad (15)$$

$$g_{i,j,w} \geq x_{i,w} - u_{i,j,w}, \forall w \in W, \forall i \in I, \forall j \in J_i \quad (16)$$

$$g_{i,j,w} \leq 2 - x_{i,w} - u_{i,j,w}, \forall w \in W, \forall i \in I, \forall j \in J_i \quad (17)$$

$$B \cdot k_{ij} \geq \sum_{w \in W} g_{i,j,w}, \forall i \in I, \forall j \in J_i \quad (18)$$

$$k_{i,j} \leq \sum_{w \in W} g_{i,j,w}, \forall i \in I, \forall j \in J_i \quad (19)$$

$$B \cdot r_w \geq \sum_{i \in I} \sum_{j \in J_i} g_{i,j,w}, \forall w \in W \quad (20)$$

$$r_w \leq \sum_{i \in I} \sum_{j \in J_i} g_{i,j,w}, \forall w \in W \quad (21)$$

$$B \cdot e \geq \sum_{i \in I} \sum_{j \in J_i} k_{i,j} \quad (22)$$

$$e \leq \sum_{i \in I} \sum_{j \in J_i} k_{i,j} \quad (23)$$

Capacity constraint of a DU is split in two parts: number of PHYcells (Eqn. (4)) and capacity of UPs (Eqn. (5)). Total redirected traffic cannot exceed the bandwidth of the switch (Eqn. (6)). Eqn. (7) ensures that every cell is served by a wavelength. Eqn. (8) ensures that every user in a cell is served by a wavelength. The network resources must be enough to accommodate total traffic demand in RAN so we can minimize the active resources. For a DU or LC, as long as it serves any

traffic, it must be active and thus it consumes energy (Eqns. (9)-(12)). The activity of an LC will lead to the activity of its supporting DU because the PHYcell of a cell must be set up in this DU. But the activity of a DU might not necessarily lead to the activity of its associated LC as there could be scenarios where the DU is only used as a dedicated container (with no PHYcell) for redirected UPs, and thus its LC is inactive (Eqn. (13)). Eqns. (14)-(17) enforce XOR between $u_{i,j,w}$ and $x_{i,w}$ to generate an auxiliary variable $g_{i,j,w}$, which checks whether a UP is redirected (Eqns. (18)-(19)) and whether a DU participates in UP redirection (Eqns. (20)-(21)). A DU participates in UP redirection if it sends UP to other DU or accommodates UP redirected to it. Existence of redirected UP indicates the activity of the switch (Eqns. (22)-(23)).

C. Complexity Analysis

Number of variables is $O(|W||I|N)$, where $|W|$ is number of wavelengths, $|I|$ is number of cells, and N is number of users. Number of constraints is also $O(|W||I|N)$. The 1D-BP problem has been proved to be a combinational NP-hard problem, thus our even harder S-VF and D-VF problems are also NP-hard. Thus, we resort to heuristic method for quickly solving the problems for large instances.

V. VBS FORMATION SCHEMES

A. Static VBS Formation Scheme

We propose the static VBS formation (S-VF) scheme, which contains two algorithms, DU packing (DUP) and Greedy Load-Shedding (GLS). DUP is designed to pack all cells into a theoretical minimum number of DUs, $\sum_{k=1}^N C_k \cdot load / C_{UP}$, by temporarily allowing DU to be overloaded. GLS is designed to rebalance overloaded DUs by PHYcell shifting and UP redirection. Our preliminary work [16] on this topic has provided the details of these two algorithms that we assume as given subroutine S_1 and S_2 in the following.

Fig. 3 presents the flowchart of S-VF. Initially, all network resources are inactive, and we are given the traffic profile over a certain period of time, based on which we form VBS for each cell. First, a PHYcell needs to be set up for each cell with users to be served. DUP is designed for the assignment of PHYcells to DUs with the constraint of DU's PHYcell capacity. Then, we estimate the DU load by summing the traffic demands of users that will be served by the DU. If any DU will be overloaded, we rebalance it by calling GLS, which rebalances overloaded DUs through PHYcell shifting. After the improvement of PHYcell assignment, we assign to each cell the wavelength of the LC connected to the DU that hosts the PHYcell set up for this cell. And for each user in the cell, we set up the UP in the hosting DU. After the UP assignment, there may exist overloaded DUs, which can be further rebalanced through UP redirection enabled by GLS. Finally, if DUs cannot be rebalanced any more, we activate a new DU as a dedicated container, in which no PHYcell is set up, for redirected UPs from overloaded DUs.

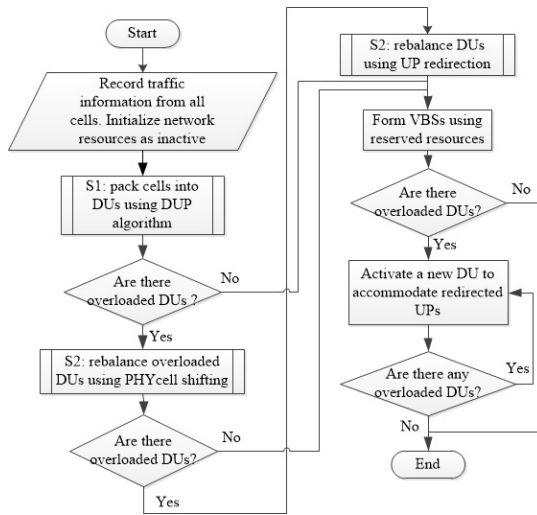


Fig. 3. Flowchart of Static VBS formation.

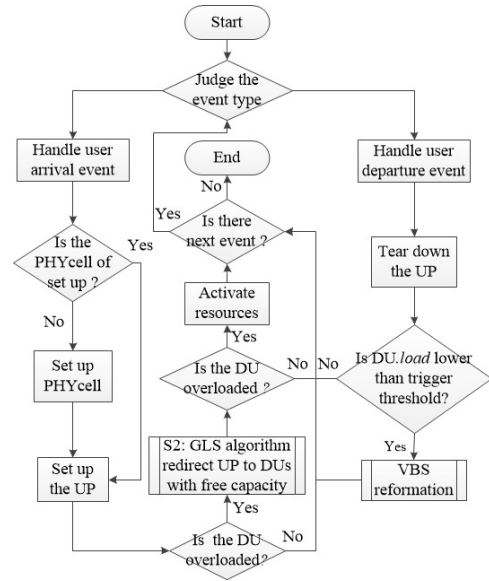


Fig. 4. Flowchart of Dynamic VBS formation.

B. Dynamic VBS Formation Schemes

Now, we propose a naive D-VF (N-DVF) scheme and a sophisticated D-VF (S-DVF) scheme respectively.

The basic idea is to use only “just-enough” resources to support current traffic demands. Unlike static case, where traffic loads for a period are given beforehand and remain unchanged, in dynamic case, traffic loads are not known a-priori, or are subject to statistical uncertainty. Now, when a connection request arrives, the centralized controller needs to make an online decision for how to provision it immediately (how and where to set up the UP?). For example, UP needs to be redirected if its setting causes the overload of its hosting DU, which raises another question: which DU should we redirect the UP to? But UP redirection might degrade the quality of service (QoS) for delay-sensitive services, e.g., video. When a connection departs, the controller needs to tear down the UP to reclaim the network resources, and decide whether and how a VF should be conducted if current VBS configurations are not energy efficient.

1) *Naive D-VF (N-DVF) Scheme:* We present a naive D-VF (N-DVF) scheme, which redirects UP to any DU with remaining capacity and triggers a VF when the last UP is torn down in a DU. As shown in Fig. 4, when a connection r arrives, we detect whether the PHYcell is set up for the cell that r comes from. If not, we should first set up the PHYcell in any DU that has extra PHYcell capacity. Then, we set up the UP for r in the DU. If the setting of r 's UP causes the overload of the hosting DU, then we redirect some UP from the hosting DU to DU with load less than C_{UP} using GLS. If the load of the hosting DU cannot be reduced below C_{UP} , we have to activate a new DU as a dedicated DU that contains no PHYcell to contain the extra UPs from the overloaded DU. When a connection r' expires, we tear down its UP and reclaim the processing resource. If r' is the last request that was served, we conduct a VF using S-VF based on current traffic loads so that we can save energy by shutting down the unnecessary DU, LC, and wavelength. However, N-DVF might be suboptimal as it incurs a large amount of redirected traffic, which can cause

delay for services. The delay of UP redirection comes from decision making and redirection procedure. If we redirect the UP to a random DU that has remaining capacity, the target DU might approach its capacity limit, and not accommodate the incoming UPs that are supposed to be set up in it. Thus, the incoming UPs will be forced to further redirect to other DUs when they arrive, which may lead to a *chain reaction* of UP redirection.

2) *Sophisticated D-VF (S-DVF) Scheme:* In S-DVF scheme, instead of randomly redirecting UP and using capacity of a DU, we reserve some capacities in a DU for incoming traffic. Also, instead of conducting a VF passively, we proactively trigger a VF according to current/historical traffic pattern.

3) *Capacity Reservation:* We reserve some capacity of a DU for incoming traffic whose UPs are supposed to be set up in the DU that contains the PHYcell for the cell that the traffic comes from. Note that capacity reservation can effectively alleviate the *chain reaction* of UP redirection. We divide the capacity of a DU into three sections, ‘occupied’, ‘reserved’, and ‘free’, where ‘occupied’ indicates the amount of baseband processing capabilities of a DU that are occupied by existing non-redirected UPs (UPs that locate in the same DU with its PHYcell), ‘reserved’ means that the part of baseband processing capabilities of the DU are reserved for incoming non-redirected UPs, and ‘free’ means that the part of baseband processing capabilities of the DU can be used to accommodate the redirected UPs (note that non-redirected UPs are allowed to occupy not only the ‘reserved’ section but also ‘free’ section when necessary). When a UP needs to be redirected, we pick the DU with the most amount of ‘free’ capacity. If there is no DU with ‘free’ capacity, then we activate a new DU as a dedicated container for redirected traffic and set its initial ‘free’ capacity as C_{UP} , because there is no PHYcell set up in it and thus no incoming non-redirected UP.

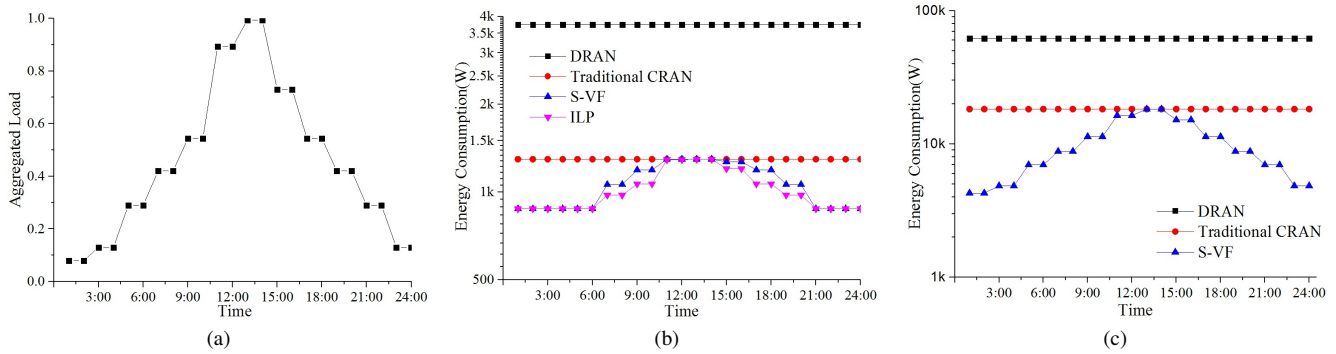


Fig. 5. (a) Daily traffic profile in typical business area. (b) Energy consumptions of different architectures in small network. (c) Energy consumptions of different architectures in large network.

4) *Proactive VF*: We set a trigger threshold for VBS reformation in terms of DU's load. When one or more DUs have loads less than the threshold, it indicates that the traffic demands in RAN are decreasing and VF is needed to adjust VBS configurations for more energy efficiency. But, if the threshold is too high, statistical uncertainty of traffic may frequently drop the DU's load below it, which may cause unnecessary VFs. If the threshold is too low, VF is difficult to be triggered, leaving VBS configurations in an energy-inefficient status for a long time. Therefore, a good choice of threshold is to make it adaptive to historical traffic pattern.

Traffic load in RAN is highly dynamic and has strong diurnal pattern [22], alternating between peak and idle durations, while traffic demands are stable over short term (e.g., same time of consecutive days). So, we must adjust the amount of free capacity and trigger threshold during a day before the traffic pattern changes in CRAN. We set the free capacity (F_t) and trigger threshold (T_t) of a DU as follows:

$$F_t = \begin{cases} 0, & V_t(1+p) \geq C_{UP} \\ C_{UP} - (1+p)V_t, & otherwise \end{cases} \quad (24)$$

$$T_t = \begin{cases} 0, & p \geq 1 \\ (1-p)V_t, & otherwise \end{cases} \quad (25)$$

where V_t is average traffic load of a DU at time t , $V_t = \frac{\lambda_t}{\mu_t} b_t$, b_t is the average bandwidth demand of a connection request, λ_t is the arrival rate at time t , and μ_t is the departure rate at time t . We introduce the trigger knob (p value) in these formulas as a network-operator-specified parameter. The increasing of the p value decreases the amount of initial 'free' capacity of a DU, and decreases the trigger threshold of VF. We can set up different values for F_t and L_t during a day by adjusting the p value or values of arrival rate, departure rate, or average bandwidth demand according to local current/historical data.

VI. NUMERICAL EVALUATION

A. Static Case

We consider CRAN topologies in two scales: small network with 6 DUs and 6 cells; and large network with 100 DUs and 100 cells. In static case, traffic loads in each period are given a-priori and only change across periods, and VF is done

once per period. As shown in Fig. 5a, we consider a typical operational day in access networks in a business area [25]. The daily 24 hours are slotted into 12 periods, each of 2-hour length. Users have busy hours from 10:00 to 17:00 and traffic loads reach the peak at 13:00-14:00. Traffic loads are normalized (aggregated load), with respect to the maximum load (M) in RAN during a day. Capacity of a wavelength is 10 Gbps. The traffic-dependent CPRI rate per user is uniformly and independently drawn from [100 Mbps, 1 Gbps] dependent on the traffic load, and the fixed CPRI rate per cell is the ratio between the wavelength capacity and the C_{PH} value³. Summary of energy modeling parameters and their values are listed in Table I. We set C_{UP} and C_{PH} as adjustable parameters to study their impacts on energy consumption.

1) *Small Network*: We first fix C_{UP} and C_{PH} to 5 Gbps and 4, respectively, and study energy consumption performances of different architectures.

In Fig. 5b, we compare energy consumptions during a day among the three architectures: CRAN with VF (including ILP and S-VF), traditional CRAN, and DRAN. DRAN consumes the most energy due to the stand-alone housing of DUs at cell sites. Traditional CRAN consumes less energy, thanks to the pooling of DUs, but without VF, all DUs need to remain active all the time, which leads to energy waste. CRAN with VF achieves significant additional energy saving, thanks to pooling of DUs and forming VBSs (sharing network resources and shutting down unused/idle ones) based on the variation of traffic load during a day.

In Fig. 5b, we see that the performance of S-VF achieves similar energy savings to ILP results and thus is benchmarked over this small network. During busy hours (early afternoon) and idle hours (early morning and night), S-VF scheme achieves the same energy consumption with ILP, which shows that S-VF is able to obtain optimal energy efficiency under stable peak and idle periods. But when traffic loads are increasing or decreasing, S-VF achieves suboptimal performance because it has a tradeoff between performance and time complexity.

2) *Large Network*: The large network used for our evaluation has 100 DUs and 100 cells (other parameters are same as

³Here, we assume fixed CPRI rate for cell processing, but traffic-dependent CPRI rate for user processing. But note that there are various ongoing projects for making CPRI rate per cell traffic dependent in [26], [27].

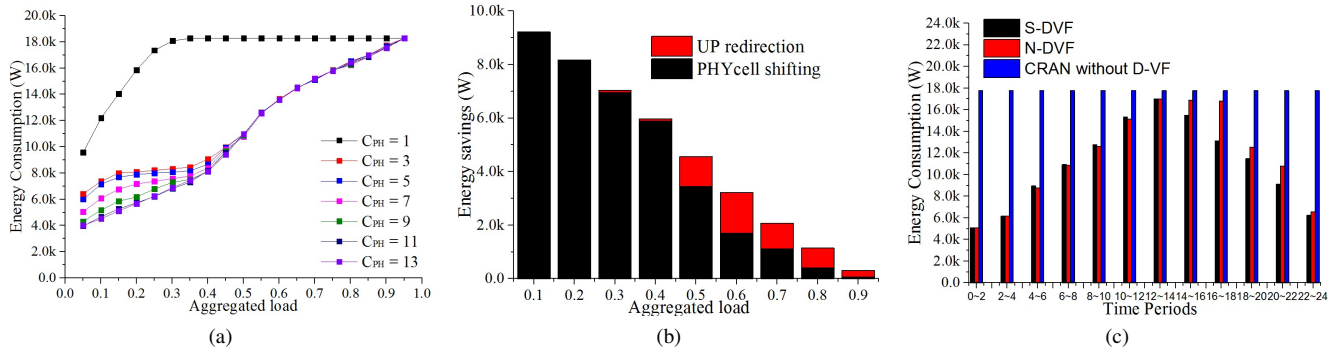


Fig. 6. (a) Energy consumptions of S-VF scheme for different C_{PH} s. (b) Energy savings of UP redirection and PHYcell shifting of S-VF scheme. (c) Energy consumptions of CRAN with two D-VF schemes and CRAN without D-VF.

in the small network). We solve the problem only by heuristic as the ILP cannot solve this case in reasonable time. In Fig. 5c, we set $C_{UP} = 5$ Gbps and $C_{PH} = 10$. Our proposed architecture with S-VF achieves even more energy savings in the large network, namely 46.1% and 84.1%, on average, compared with traditional CRAN and DRAN, respectively. This is because, compared to small network, large network has more DUs pooled together, shares common housing facilities, and has more opportunities to consolidate extra DUs.

In Fig. 6a, we first fix $C_{UP} = 5$ Gbps, and plot energy usage for different C_{PH} values as we increase the aggregated load. When aggregated load is larger than circa 0.5, all C_{PH} values achieve the same energy consumption because, as the aggregated load increases beyond 0.5, loads in cells become so high that rare DUs can be shared, and thus the C_{PH} makes less difference on energy consumption. UP redirection will balance the load among DUs regardless of the C_{PH} . We find that, when C_{PH} value is larger than 10, further increasing C_{PH} value makes little difference in energy saving, because we assume that the DU's UP capacity is constrained, so multiplexing gain is limited even if we allow a large number of PHYcells to be accommodated by a single wavelength/DU. So when CPRI rate per cell is larger than 1 Gbps, the smaller the CPRI rate per cell is, the more energy saving can be achieved. We also find that putting a constraint on the number of DUs, over which UPs of a single cell can distribute, will reduce the energy savings (limiting the UP redirection). And our proposed VF can consolidate UPs to the minimum number of DUs when they are distributed dispersively in the DU cloud.

In Fig. 6b, we show the energy savings of our proposed scheme compared to traditional CRAN. We fix C_{PH} at 10. Each column shows two contributions to energy saving: PHYcell shifting and UP redirection. As expected, we see that energy savings increase as traffic load decreases because, when traffic load is low, we can consolidate more cells in a single DU and VPON, and thus use less active LCs, DUs, and switch ports. We also find that UP redirection tends to play a noticeable role in saving energy when traffic load is above 0.5. This is because, when traffic load is above 0.5, it is hard for DU to save energy by accommodating more than one cell (PHYcell shifting), but DUs can be still utilized efficiently, thanks to UP redirection. The energy savings due to

UP redirection will be more significant if the traffic load from a single cell is more comparable with the DU's UP capacity.

B. Dynamic Case

When duration and arrival time of traffic demands are unknown, when and how to form VBS must be decided at run time. We proposed two dynamic VF schemes, called N-DVF and S-DVF. An event-driven simulator has been developed to compare the two approaches on the large topology, with the same input settings used for the static scenario. Offered traffic follows the same daily variations shown in Fig. 5a, but during each two-hour period, traffic request arrives following a Poisson distribution and holds for a duration following a negative exponential distribution.

We first study the impact of the p value on the performance of S-DVF scheme. We plot total energy consumption, total redirected traffic, and number of VFs of S-DVF for increasing p value in Figs. 7. In Fig. 7a, we can see that the energy consumption increases for increasing p value, as (see Eqn. (24)) large p value decreases the initial 'free' capacity of a DU, but the more free capacity, the less DUs need to be activated. Fig. 7b shows that higher p values also lead to larger amount of redirected traffic. In fact, following a similar reasoning, large p values decrease the trigger threshold of VBS formation (see Eqn. (25)), and therefore, with less VBS formations, energy savings are achieved by redirecting more UPs (instead of frequently reforming VBSs). In conclusion, larger p values can effectively avoid excessive VBS formations, as shown in Fig. 7c. Thus, for the following analysis, we choose a moderate p value (0.55), that provides a good tradeoff between the amount of VBS formations in a day (only few tens in a day), and the energy savings and redirected traffic.

In Fig. 6c, we plot the energy consumptions of S-DVF and N-DVF schemes, and compare them with that of CRAN without D-VF. Significant energy savings can be achieved by D-VFs compared to CRAN without D-VF. As shown in Fig. 6c, S-DVF consumes less energy than N-DVF after 14:00, when traffic load begins to decrease. That is because, in S-DVF, VF is easier to be triggered and network resources (DUs and LCs) are consolidated more efficiently following the variation of traffic loads in RAN.

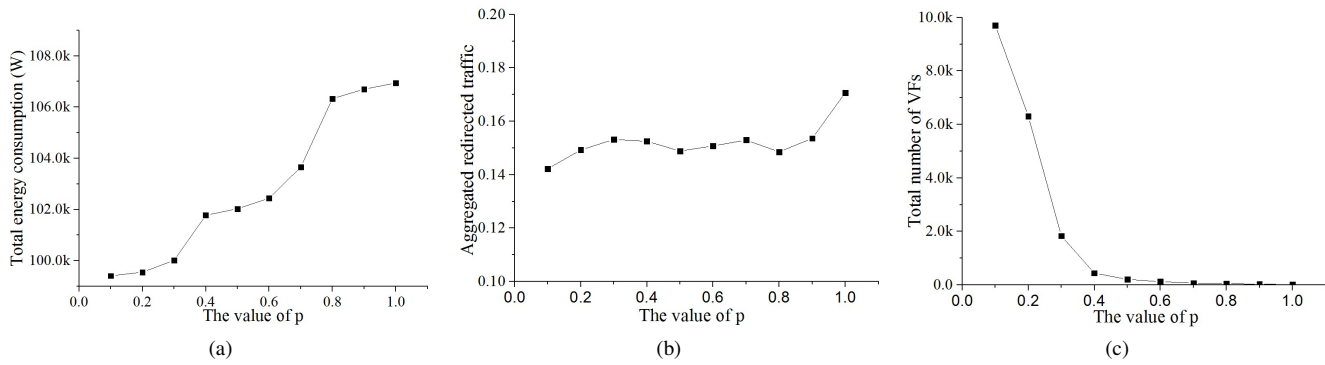


Fig. 7. (a) Total energy consumption. (b) Aggregated load of total redirected traffic. (c) Total number of VFs.

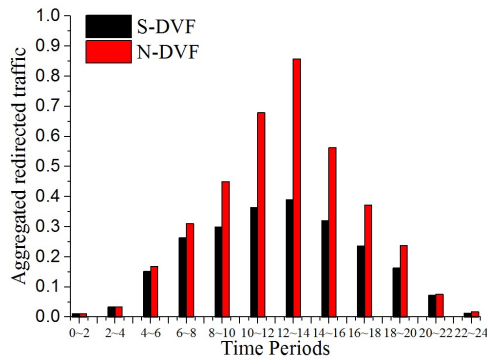


Fig. 8. Redirected traffic of S-DVF scheme and N-DVF scheme during 12 time periods of a day.

In Fig. 8, we compare the aggregated redirected traffic in S-DVF and N-DVF. S-DVF achieves a significant reduction of redirected traffic compared to N-DVF as S-DVF reserves part of DU’s capacity for incoming traffic, which avoids the *chain reaction* of UP redirection. Thus, S-DVF provides better QoS than N-DVF for delay-sensitive traffic.

We also compare the number of VFs in S-DVF and N-DVF in a day, which are 8 and 56, respectively. When p value is properly chosen, S-DVF reconfigures a much lower number of VFs during a day while still achieving relevant energy savings.

VII. CONCLUSION

We proposed an energy-efficient architecture of cloud radio access network (CRAN). Based on the architecture, we proposed a novel concept, virtual base station (VBS), which is a combination of a virtualized fronthaul link (a virtual PON) and multiple virtualized functional entities of baseband processing. By flexibly forming a VBS for each cell, network resources can be effectively shared to save energy in CRAN. We formulated the energy-efficient VBS formation (VF) optimization problem using an Integer Linear Program (ILP) with the objective of minimizing overall energy consumption in CRAN. In network planning stage, where (a forecast of) future traffic is given, we designed a fast heuristic algorithm for static VBS formation (S-VF). In traffic engineering stage, where future traffic is unknown and connection requests arrive and depart, we designed dynamic VBS formation (D-VF) schemes.

Extensive simulations showed that our proposed CRAN with S-VF and D-VF achieve significant energy savings compared to traditional distributed RAN (DRAN) and CRAN without VF. In static case, S-VF achieves near-optimal performance of energy consumption. In dynamic case, D-VF scheme showed its superiority on energy consumption and quality of service.

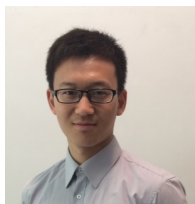
ACKNOWLEDGMENT

This work was supported by ICT R&D program of MSIP/IITP. [14-000-05-001, Smart Networking Core Technology Development].

REFERENCES

- [1] J. Baliga et al., “Energy consumption in access networks,” *Proc. OFC*, San Diego, US, 2008.
- [2] China Mobile Research Institute, “C-RAN: The road towards green RAN,” China Mobile Research Institute, 2011.
- [3] P. Rost, et al., “Cloud technologies for flexible 5G radio access networks,” *IEEE Communications*, vol. 52, no. 5, pp. 68-76, May 2014.
- [4] M. Peng, et al., “System architecture and key technologies for 5G heterogeneous cloud radio access networks,” *IEEE Network*, vol. 29, no. 2, pp. 6-14, April 2015.
- [5] T. Pfeiffer, “Next Generation Mobile Fronthaul Architectures,” *Proc. OFC*, Los Angeles, US, 2015.
- [6] S. Bhaumik et al., “CloudIQ: a Framework for Processing base stations in a data center,” *Proc. ACM Mobicom*, Istanbul, Turkey, 2012.
- [7] M. Qian et al., “Baseband Processing Units Virtualization for Cloud Radio Access Networks,” *IEEE Wireless Communications Letters*, vol. 4, no. 2, pp. 189-192, April 2015.
- [8] B. Haberland et al., “Radio Base Stations in the Cloud,” *Bell Labs Tech. Journal*, vol. 18, no. 1, pp. 129152, May 2013.
- [9] C. Liu et al., “A Novel Multi-Service Small-Cell Cloud Radio Access Network for Mobile Backhaul and Computing Based on Radio-Over-Fiber Technologies,” *IEEE/OSA Journal on Lightwave Technology*, vol. 31, no. 17, pp. 2869-2875, September 2013.
- [10] D. Iida et al., “Dynamic TWDM-PON for Mobile Radio Access Networks,” *Optics Express*, vol. 21, no. 22, pp. 26209-26218, 2013.
- [11] J. Zhang et al., “Designing energy-efficient optical line terminal for TDM passive optical networks,” *Proc. Sarnoff Symp.*, Princeton, 2011.
- [12] R. Wang et al., “Energy Saving via Dynamic Wavelength Sharing in TWDM-PON,” *IEEE J. on Selected Areas in Commun.*, vol. 32, no. 8, pp. 1566-1574, Aug. 2014.
- [13] Y. Ji et al., “All Optical Switching Networks With Energy-Efficient Technologies From Components Level to Network Level,” *IEEE J. on Selected Areas in Commun.*, vol. 32, no. 8, pp. 1600-1614, Aug. 2014.
- [14] M. Peng et al., “Energy-Efficient Resource Assignment and Power Allocation in Heterogeneous Cloud Radio Access Networks,” *IEEE Transactions on Vehicular Technology*, vol. PP, no. 99, Dec. 2014.
- [15] Z. Kong et al., “eBase: A baseband unit cluster testbed to improve energy-efficiency for cloud radio access network,” *Proc. ICC*, Budapest, 2013.

- [16] X. Wang et al., "Green Virtual Base Station in Optical-Access-Enabled Cloud-RAN," *Proc. ICC*, London, UK, 2015.
- [17] CPRI Specification V6.0; August 30, 2013; <http://www.cpri.info>.
- [18] A. Pizzinat et al., "Things You Should Know About Fronthaul," *Journal on Lightwave Technology*, vol. 33, pp. 1077-1083, March 2015.
- [19] N. Carapellese et al., "Energy Efficient BaseBand Units (BBU) Placement in a Fixed/Mobile Converged WDM Aggregation Network," *IEEE J. on Selected Areas in Commun.*, vol. 32, pp. 1542-1551, Aug. 2014.
- [20] iCIRRUS Project. <http://www.icirrus-5gnet.eu/project/>.
- [21] Y. Yang, "Understanding Switch Latency White Paper," Cisco. CA.
- [22] A. Chandrakasan et al., "Low-power CMOS digital design," *IEEE Trans. on Solid-State Circuits*, vol. 27, no. 4, pp. 473-484, Apr. 1992.
- [23] A. Dixit, et al., "ONU power saving modes in next generation optical access networks: progress, efficiency and challenges," *Optics Express*, vol. 20, no. 26, pp. 52-63, Nov. 2012.
- [24] A. Conte et al., "Power consumption of base stations," *Trend Plenary Meeting for 7th Framework Programme*, Alcatel-Lucent Bell Labs, 2012.
- [25] P. Chunyi, "Traffic Driven Power Saving in Operational 3G Cellular Networks," *Proc. ACM Mobicom*, US, 2011.
- [26] NGMN White paper, "Future Study On Critical C-RAN Technologies," NGMN, March, 2015.
- [27] IEEE 1904 Access Networks Working Group, "Some Views on Next Generation Radio Interface," Feb. 2015.



Xinbo Wang Xinbo Wang is currently a Ph.D. candidate in computer science at the University of California, Davis. He received the B.Eng. in Telecommunication Engineering from Beijing University of Posts and Telecommunication, Beijing, China, in 2013, and the M.S. in computer science from University of California, Davis, USA in 2015. His research interests include cloud radio access network and optical transport network for 5G.



Saigopal Thota Saigopal Thota currently works as a software engineer in Cablevision. Saigopal received his PhD degree in Computer Science from the University of California, Davis in 2014 and B.Tech. degree from Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, India, in 2009. His research interests include cloud-based service architectures for heterogeneous multi-device collaboration and hybrid optical access networks.



Hwan Seok Chung Hwan Seok Chung received the Ph.D. degree in electronics engineering from Korea Advanced Institute of Science and Technology (KAIST) in 2003. In 2003, he was a Postdoctoral Research Associate with KAIST, where he worked on hybrid CWDM/DWDM system for metro area network. From 2004 to 2005, he was with KDDI R&D laboratories Inc., Saitama, Japan, and had engaged in research on wavelength converter and regenerator. Since 2005, he has been with Electronics and Telecommunication Research Institute (ETRI), Daejeon, where he is currently a director of optical access research section. His current research interests include mobile fronthaul, high-speed PON, and modulation format. Dr. Chung served as a technical committee member of OFC, OECC, COIN, ICOCON, and Photonic West. He was the recipient of the Best Paper Awards from the Optoelectronics and Communications Conference (OECC) in 2000 and 2003 as well as ETRI in 2011 and 2012. Dr. Chung is senior member of IEEE.



Han-Hyub Lee Han-Hyub Lee received his BS, MS, and Ph.D. degrees in physics from Chungnam National University, Daejeon, Rep. of Korea, in 1999, 2001, and 2005, respectively. His doctoral research included the application of a Raman fiber amplifier and gain-clamped SOA for WDM systems. From 2006 to 2007, he was a postdoctoral researcher at AT&T Laboratory, Middletown, NJ, USA, where he worked on extended WDM/TDM hybrid PONs using a wideband optical amplifier. In 2007, he joined ETRI as a senior researcher of the Department of Optical Internet Research. He has worked on optical access networks and has contributed to the development of international standardizations. He is a member of the IEEE.



Soomyung Park Soomyung Park was born in Seoul, Korea. in 1968. He received the B.S. degree in computer science engineering from the University of Dankuk, in 1990, and M.S. and Ph.D. Degrees in computer engineering from the University of Konkuk, in 1992 and 1999, respectively. Since May 2000, he has been with the Department of Communication and Internet, ETRI. His current research interests include the transport (MPLS-TP packet transport network and Optical Transport Network) SDN (Software Defined Networking) area and open source based SDN projects like OpenDaylight, ONOS, etc.

Biswanath Mukherjee Biswanath Mukherjee is a Distinguished Professor at University of California, Davis. For a photo and bio, please see p. 1574, vol. 32, no. 8, August 2014, of this journal.

Massimo Tornatore Massimo Tornatore is an Associate Professor at Politecnico di Milano, Italy. For a photo and bio, please see p. 1551, vol. 32, no. 8, August 2014, of this journal.