Can Fine-Grained Functional Split Benefit to the Converged Optical-Wireless Access Networks in 5G and Beyond?

Yuming Xiao, *Graduate Student Member, IEEE*, Jiawei Zhang, *Member, IEEE*, and Yuefeng Ji, *Senior Member, IEEE*

***Abstract*—The centralized radio access network (C-RAN) is an effective architecture to promote CAPEX/OPEX reduction and cell cooperation derived from its centralized baseband processing. However, there is a contradiction between centralization gain and transport resource saving, which hinders the vision of a resource-efficient and cost-effective RAN deployment. Advanced RAN architectures with functional splits are then introduced to cope with this challenge. Distinguished with other studies, we are intended to investigate whether a fine-grained functional split architecture could benefit to the RAN evolution, and how it impacts on the converged optical-wireless access networks. To this end, we establish a quantitative model to analyze the performance of this architecture. With the fine-grained split, baseband unit (BBU) is divided into a set of fine-grained units (FU) to be placed in desired processing pools (PP) as a service chain. To analyze the placement performance, we propose a mixed-integer linear programming model (MILP) considering the PP selection, routing, wavelength and bandwidth assignment, as well as latency control to minimize the number of PPs, bandwidth, latency, and functions deployment cost. We compare its performance with other two coarse-grained split architectures, i.e., SBBU (adopt low-PHY split like BBU in 4G) and recently emerged DU-CU in both small-scale and large-scale network scenarios. Our analyses provide insights into the modeling and design of efficient converged optical-wireless access networks in 5G and beyond.**

***Index Terms*—Baseband processing placement, functional split, converged optical-wireless networks.**

I. Introduction

FIFTH-GENERATION (5G) mobile communication has launched a high-speed, intelligent, and interconnected era, which is intended to deliver a 1000-fold higher data rate, reduce round-trip latency, and support more connected smart devices than 4G [1], [2]. Considerable challenges have emerged to stimulate new design principles on radio access networks (RANs) [3], [4]. Centralized RAN (C-RAN) architecture, in which the legacy base station (BS) is disaggregated into remote radio unit (RRU) and baseband unit (BBU), is widely used in 4G for the CAPEX/OPEX reduction and cell cooperation. The RRU, interfaced with antennas, is responsible for the transmission and reception of radio signals, analog/digital (A/D) conversion and power amplification. The BBU performs the digital baseband signal processing and radio resource scheduling, which is centralized in one or several common processing pools (PPs) for sharing computational resources and housing/cooling facilities. The optical transport network is considered as an outstanding media to connect RRUs and PPs with a load-independent common public radio interface.

In the upcoming 5G/B5G era, the advanced wireless technologies and unprecedented service experience will be incorporated. A contradiction is then raised between baseband processing centralization and transport resource saving that prevents from a resource-efficient and cost-effective RAN deployment. For example, fronthaul suffers from the large bandwidth requirement to transmit raw in-phase and quadrature (I/Q) samples under the continuous expansion of the wireless spectrum and antenna scale. If BBUs are highly centralized into remote PPs, then 100-fold or even 1000-fold transport bandwidth will be required than 4G. If BBUs are distributed to the RRU side for bandwidth saving, then more edge PPs will be built with extra expenditure and power consumption. In order to deploy a resource-efficient and cost-effective network, advanced RAN architectures with functional splits are introduced.

The functional split is to re-define the BBU into different function entities, which has been discussed both in academic and industrial fields. Next-generation RAN (NG-RAN) architecture is recognized as the solution for 5G, in which BBU is split into two new entities named distributed unit (DU) and centralized unit (CU) [5]. The large-bandwidth and latency-sensitive BBU functions below packet data converged protocol layer (PDCP) are provided in DU, while functions above PDCP are provided in CU. Two split options are adopted in NG-RAN to make it serve as a three-tier architecture. Recently, with the development of network function virtualization paradigm, network functionality is split up and modularized into multiple building blocks that can be chained

Manuscript received November 20, 2019; revised March 25, 2020; accepted May 13, 2020. Date of publication May 20, 2020; date of current version September 9, 2020. This work was supported by the National Key R&D Program of China (No. 2018YFB1800802), the National Nature Science Foundation of China Projects (No. 61871051), the Beijing Natural Science Foundation (No. 4192039), the fund of State Key Laboratory of Advanced Optical Communication Systems and Networks, China, No. 2019GZKF5, and BUPT Excellent Ph.D. Students Foundation (No. CX2019222). The associate editor coordinating the review of This article and approving it for publication was R. Riggio. *(Corresponding authors: Jiawei Zhang; Yuefeng Ji.)*

The authors are with the State Key Laboratory of Information Photonics and Optical Communications, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: zjw@bupt.edu.cn; jyf@bupt.edu.cn).

Digital Object Identifier 10.1109/TNSM.2020.2995844

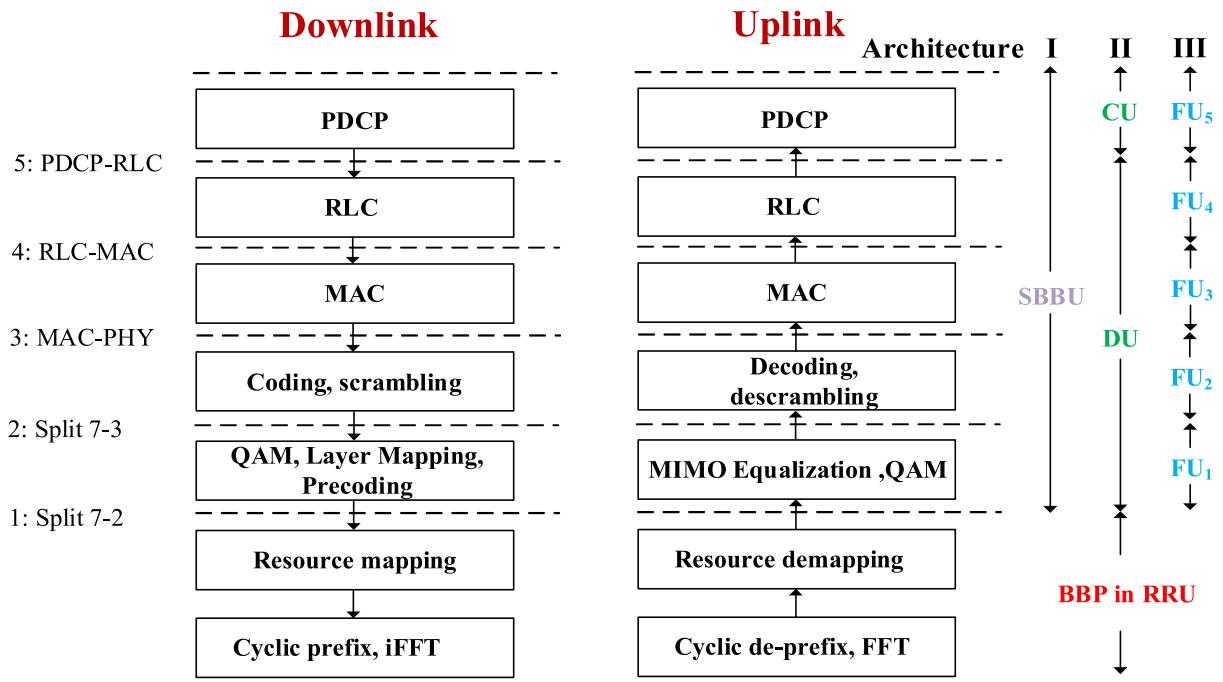

Fig. 1. Three split architectures where split options follow the 3GPP specification [5].

together for specific purposes. Inspired by this point, whether the fine-grained functional split can benefit to the RAN in 5G and beyond deserves the further discussion.

In this study, we are intended to discuss whether a fine-grained split architecture can deeply help RAN for cost saving. With the fine-grained split, BBU is split into a set of fine-grained units (FU) to be placed into desired PPs as a service chain. In this paper, we focus on providing a quantitative model to analyze the performance of this architecture. We then formalize the FU placement problem as a multi-objective mixed-integer linear programming model considering the PP selection, routing, wavelength and bandwidth assignment, as well as latency control. We aim to minimize the number of used PPs (i.e., centralization gain), bandwidth consumption, latency, and network deployment cost. To evaluate the fine-grained split performance, we compare FU architecture with other two coarse-grained split architectures, i.e., RRU-SBBU and RRU-DU-CU. In three architectures, we adopt functional split options proposed by 3GPP [6]. The low-physical layer (low-PHY) functions are located with RRU, whereas remaining functions are provided in PPs. As shown in Fig. 1, 1) Architecture I: RRU-SBBU (SBBU, i.e., Split BBU) only adopts the option *Split 7-2* to divide BBU into two parts. The remaining functions are then combined as an SBBU, which is like BBU in 4G; 2) Architecture II: RRU-DU-CU adopts options *Split 7-2* and *Split PDCP-RLC* to divide BBU into three parts. The functions of high-PHY, media access control (MAC) and radio link control (RLC) layers are provided in DU, while functions above PDCP are provided in CU; 3) Architecture III: RRU-FU adopts all available split options to divide the remaining functions into five FUs. A generalized interface (GI) between any adjacent FUs should be introduced to support any possible data rate. We compare the baseband processing placement of three split architectures in both small-scale and large-scale networks with our proposed MILP. Our analyses can provide insights into the modeling and design of efficient converged optical-wireless access networks in 5G and beyond.

The rest of this paper is organized as follows. We report the current state of research under different split architectures in Section II, where our contributions are also clarified. Sections III and IV introduce the functional split-enabled network architecture and model for computational complexity and transport bandwidth, respectively. Section V describes the FU placement scheme and latency model, while Section VI proposes the MILP model. Section VII presents the simulation results to compare the performance of three split architectures. Section VIII draws the conclusion and provides a comprehensive evaluation on the fine-grained split.

II. Related Works and Contributions

A. Related Works

Several relevant studies have been conducted on the deployment optimization for 5G/B5G RAN, which can be categorized into three aspects: 1) designing an effective baseband processing (BBP) placement scheme based on non-split (i.e., BBU) or specified split (i.e., DU-CU) architectures; 2) finding the optimal functional split point between RRU and BBU, which is shown as a dual-site processing manner; 3) focusing on optical-enabled flexible networking.

For BBU architecture, [7] investigated the BBU placement optimization problem for C-RAN over a WDM aggregation network, which is evaluated under the scenarios of 1) converged fixed and mobile traffic, 2) OTN and overlay

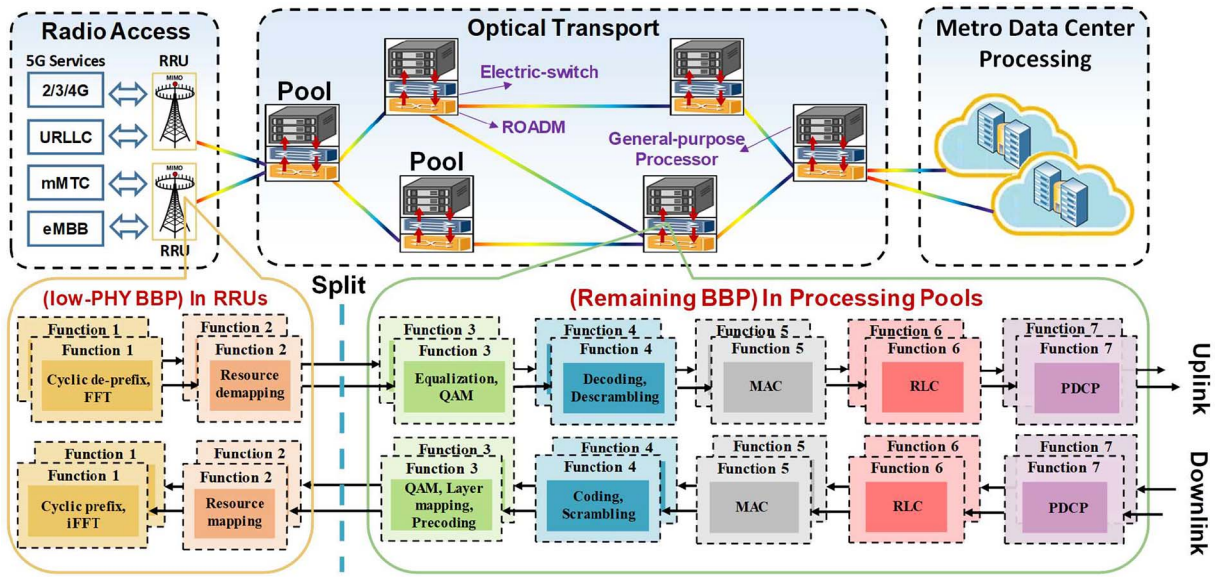

Fig. 2. Illustration of functional split-enabled network architecture.

fronthaul options, and 3) joint placement of BBU and electronic switches. In [8], BBU pool allocation and selection were jointly optimized to maximize wireless traffic capacity with the minimum wavelength consumption. Reference [9] proposed an adaptive BBU placement strategy for C-RAN under a dynamic traffic case. The simulation result showed that adaptive placement can achieve a balance between BBU centralization and traffic blocking probability. The above works optimized the BBU placement for resource saving without the functional split considered.

For DU-CU architecture, [10] optimized the DU-CU deployment using a mixed-integer quadratic programming model considering the cost of DUCU pools and links. Reference [11] focused on DU-CU placement in a slice-enabled disaggregated RAN. In our previous study [12], efficient placement of DU-CU was investigated to jointly optimize the computational and bandwidth resources. The above studies were based on the novel NG-RAN architecture but ignored other potential split options.

The dual-site processing manner is also a potential solution for resource-saving and latency satisfaction. Reference [13] investigated the selection of demarcation point between centralized and distributed units to jointly minimize the inter-cell interference and fronthaul bandwidth utilization. Reference [14] evaluated the optimal split option for a BS to minimize the total cost of ownership for Hybrid-RAN. The numerical results showed that the optimal split depended on BS configuration and data transmission direction. Reference [15] proposed a novel concept of "Dis-aggregated RAN" on converged optical-wireless networks. Energy efficiency was compared among D-RAN, C-RAN, and Dis-Aggregated RAN, in which the flexible split in the physical layer is considered. In the above studies, BBU was flexibly divided into two parts and then placed in RRU and one PP respectively. However, this solution will increase the single RRU cost to hinder the cell densification in 5G/B5G.

For flexible optical networking, a reconfigurable fronthaul network architecture to support "any-RRH to any-BBU" connection was essential. References [16], [17] designed this architecture and verified it with SDN-enabled orchestration to reduce the data exchange between different BBUs for CoMP services in coordinated radio and optical networks.

B. Contributions

Our contributions and differences from the above mentioned studies are summarized as follows.

1) This work investigates whether fine-grained split architecture can benefit to the RAN evolution and how it impacts on the converged optical-wireless access networks. A quantitative model to analyze the performance of the architecture is established.

2) This work compares the fine-grained split with two existing coarse-grained split architectures in terms of the centralization gain, bandwidth consumption, latency, and network deployment cost. We also analyze the influence of the split granularity on resource saving.

3) This work proposes a MILP model to evaluate the performance of different split architectures considering PP selection, routing, wavelength and bandwidth assignment, as well as latency control. The evaluation results can provide an insight into achieving a resource-efficient and cost-effective RAN architecture.

III. Functional Split-Enabled Network Architecture

We consider a flexible transport network that supports the multipoint-to-multipoint connections (i.e., any-RRU to any-PP) in Fig. 2. Several fibers are provided on each link, and each of them comprises multiple wavelengths. Each node is equipped with an electronic switch (E-switch) and a reconfigurable add/drop multiplexer (ROADM), which are responsible

for the traffic switching on electronic and optical domains, respectively. PP also comprises several general-purpose processors (GPPs). The GPPs enable the virtualization of baseband functions in virtual machines or containers to facilitate the efficient sharing of computational resources and reduce the processor cost [18], [19]. In this study, the low physical-layer functions are located with RRU, whereas FUs (Functions 3$\sim$7 in Fig. 2) are flexibly placed in PPs determined by MILP model. The RRU is connected with one PP (i.e., local PP) through the direct connection of fibers. The user traffic from one RRU can be processed in the local PP or any other PPs constrained to the computational and bandwidth capacity and latency requirement. All traffic flows are assumed to be aggregated into the DC node for 5G core and content processing.

The multilayer ROADM-based transport network possesses several advantages. First, this network is designed with multiple switching granularity and elastic resource allocation. Second, the flexible BBP placement under different split architectures can be supported to satisfy diversified service requirements. Third, the re-configuration of lightpaths and BBP locations is provided to guarantee the network resiliency. Fourth, this network can afford shared network infrastructures for fronthaul, midhaul, and backhaul for resource efficiency. Finally, it can perform as a potential candidate for the fixed/mobile converged network.

IV. Functional Split Model

In this section, we present a detailed description of the computational complexity and bandwidth model for each functional split. The formulas for upstream are described as follows.

A. Description of Functional Split

Based on 3GPP specifications, low-PHY functions are disaggregated from BBU to the RRU side, including the cyclic de-prefix, fast Fourier transform (FFT), and resource de-mapping. The remaining functions of high-PHY, MAC, RLC, and PDCP layers are provided in PPs. In the following part, we provide a model for the computational and bandwidth demands of each FU. The demand for SBBU (FU1-FU5), DU (FU1-FU4), and CU (FU5) can also be obtained through the model. The baseband functions in each FU are detailed as follows [20]–[22]:

- *FU1:* For upstream, FU1 comprises functions of channel estimation, MIMO equalization, and de-modulation (QAM). For downstream, functions of modulation, layer mapping, and precoding are included.
- *FU2:* For upstream, FU2 comprises the functions of descrambling, rate de-matching, and channel decoding. For downstream, scrambling, rate matching, and channel coding are contained.
- *FU3:* FU3 is responsible for MAC processing, including the functions of hybrid automatic repeat request (HARQ), mapping between logical and transmission channels, multiplexing/de-multiplexing MAC service data unit or protocol data unit (SDUs, PDUs) to/from PHY, user priority management, and logical channel priority management.
- *FU4:* FU4 completes the RLC layer processing, including the segmentation or reassembly of RLC SDUs, error correction through ARQ, and RLC re-establishment.
- *FU5:* FU5 completes the processing in the PDCP layer, including the maintenance functions of PDCP sequence number, IP packet header compression or decompression, ciphering or deciphering, and integrity protection and verification in the PDCP layer.

B. Model for Computational Complexity

To quantitatively analyze the computational complexity of FUs, a universal model derived from [23] is presented. The digital baseband processing is modeled based on estimated complexity in GOPS (Giga operations per second), which denotes the computational capacity needed in GPPs. The resource block (RB), number of antennas (A), number of MIMO layers (L), modulation bits (M), and coding rate (C) are included. The M and C are determined from modulation and coding scheme (MCS) in [24].

$$C_{FUi}[GOPS] = \alpha_i \cdot \left(3 \cdot A + A^2 + \frac{1}{3} \cdot M \cdot C \cdot L\right) \cdot \frac{RB}{5} \tag{1}$$

where α_i is a factor defined in [23]. However, [23] provided a sum of computational demands for α_1 and α_2, but their respective demand is absent. We then calculate the α_1 and α_2 through the model proposed in [25], which has estimated the physical functions using the real experimental results.

C. Model for Bandwidth Requirement

The bandwidth model is provided to estimate the data rate of each split option. We use the formulas in [26] as a basis and update them with the parameters and calculation method in the latest standard specification and reference [24], [27]. The calculation formulas are presented as follows (Mbps):

$$B_{Split7-2} = N_{SYM} \cdot N_{SC} \cdot RB \cdot A \cdot BTW/1000 \tag{2}$$

$$B_{Split7-3} = N_{SYM} \cdot N_{SC} \cdot RB \cdot M \cdot L \cdot N_{LLR}/1000 \tag{3}$$

$$B_{MAC-PHY} = TBS \cdot N_{TBS}/1000 \tag{4}$$

$$B_{RLC-MAC} = \frac{TBS \cdot N_{TBS} \cdot \left(IP_{pkt} + H_{PDCP} + H_{RLC}\right)}{\left(IP_{pkt} + H_{PDCP} + H_{RLC} + H_{MAC}\right) \cdot 1000} \tag{5}$$

$$B_{PDCP-RLC} = \frac{TBS \cdot N_{TBS} \cdot \left(IP_{pkt} + IP_{PDCP}\right)}{\left(IP_{pkt} + H_{PDCP} + H_{RLC} + H_{MAC}\right) \cdot 1000} \tag{6}$$

$$B_{Backhaul} = \frac{TBS \cdot N_{TBS} \cdot IP_{pkt}}{\left(IP_{pkt} + H_{PDCP} + H_{RLC} + H_{MAC}\right) \cdot 1000} \tag{7}$$

where *TBS* represents the bit size of transport block (TB), which is calculated through the 3GPP specification [24]. The other parameters are defined as: number of I and Q bits (*BTW*, 16 + 16 bits), number of symbols per sub-frame

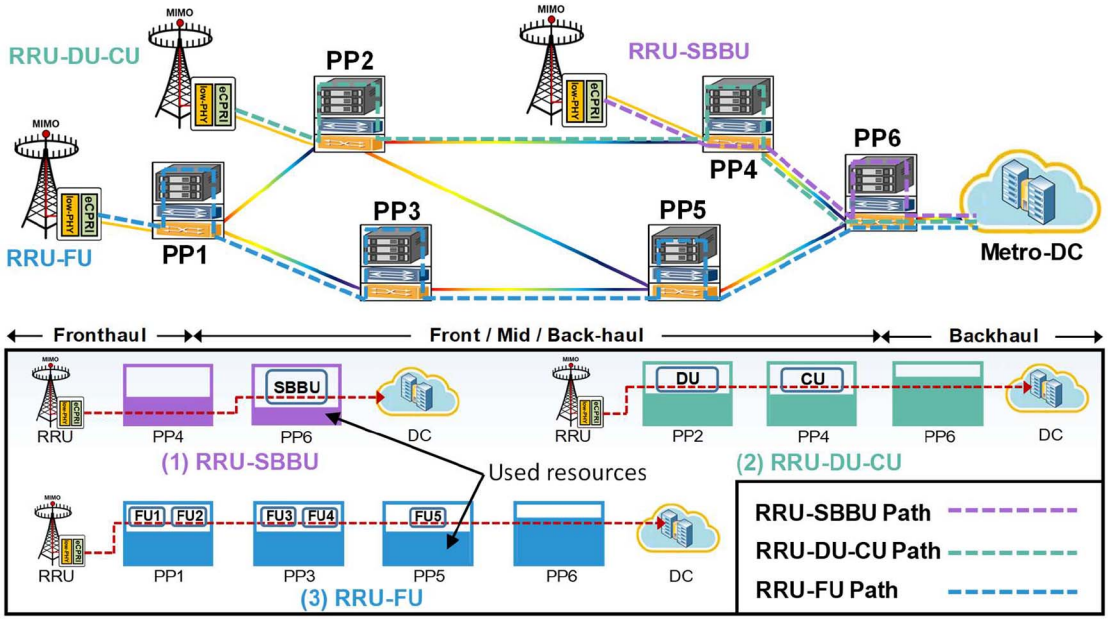

Fig. 3. Examples of BBP placement for three split architectures.

(N_{SYM}, 14), number of subcarriers per RB (N_{SC}, 12), bits of LLR (N_{LLR}, 8 bits), number of TBs per TTI (N_{TBS}, 1), IP packet size (IP_{pkt}, 1500 bytes), header size of MAC per IP packet (H_{MAC}, 2bytes,), header size of RLC per IP packet (H_{RLC}, 5bytes), PDCP header size for AM mode (H_{PDCP}, 2 bytes).

V. BBP Placement Scheme

A. Placement Scheme for Three Split Architectures

The BBP placement of three split architectures is discussed in this section. For each architecture, a complete BBP chain should be established from source RRU to DC, in which each function block should be placed once following the predefined order in Fig. 1. For instance, FUs placement cannot violate the processing order: FU1 $\rightarrow$ FU2 $\rightarrow$ FU3 $\rightarrow$ FU4 $\rightarrow$ FU5. Different split architectures also have their specific placement constraints. For RRU-SBBU, all functions (i.e., SBBU) should be placed in a common PP. For RRU-DU-CU, DU and CU can be placed together or separated into two different PPs. For RRU-FU, FUs can be placed together or separated into 2-5 different PPs.

As shown in Fig. 3, an example is provided to explain the BBP placement of three split architectures. 1) For RRU-SBBU, a chain is established with SBBU placement in PP6, in which a lightpath is established from RRU to DC passing through PP4 and PP6. The SBBU cannot be placed in PP4 because of its insufficient residual capacity for an entire SBBU processing. Thus, the remaining computational resource in PP4 is wasted. 2) For RRU-DU-CU, a chain is established with DU and CU respective placement in PP2 and PP5, in which its lightpath passes through PP2, PP5, and PP6. DU and CU require less computational resource than SBBU that enables their placement into high-load PPs. 3) For RRU-FU, a chain is established with FUs placement in three PPs of higher load, in which its lightpath passes through PP1, PP3, PP5, and PP6. RRU-FU has the highest resource efficiency because of its fine-grained split.

B. Establishment of Lightpath

The establishment of lightpath to interconnect source RRU, DC and selected PPs is necessary. There are three aspects should be considered. First, the service traffic from RRU to DC should be routed and passes through all selected PPs. Second, wavelength and bandwidth on each link should be sufficient to accommodate the RRU data. Finally, the latency should be controlled for service requirement satisfaction.

C. Latency

This latency model includes four parts:

(1) Transmission latency on fiber links which is linear with the length of fiber links (i.e., 5μs/km).

(2) Traffic switching latency in ROADM and electric switch. The ROADM is responsible for traffic switching on a wavelength basis with short latency, while electric switch supports a finer-grained switching using time slots with the latency of optical-electric-optical conversion (OEO) and electric switching (E-switching).

In Fig. 4, we discuss all situations of OEO and E-switching. The S1 $\sim$ S3 represents the service data of RRU1 $\sim$ RRU3. 1) As shown in Fig. 4 (a), switching occurs only in the optical domain named "bypass", in which traffic enters one port of ROADM device and leaves from another port without any processing in the electronic domain. This is a short-time switching that can be ignored due to the absence of OEO conversion and E-switching. 2) This situation denotes the switching in the electronic domain where BBP occurs. As shown in Fig. 4 (b), all services carried on λ 1 will experience the OEO and E-switching (e.g., for service $S2$) because $S1$ is processed in PP. 3) Fig. 4 (c) denotes the switching caused by traffic grooming. The OEO and E-switching are introduced for $\lambda 1$ and $\lambda 2$

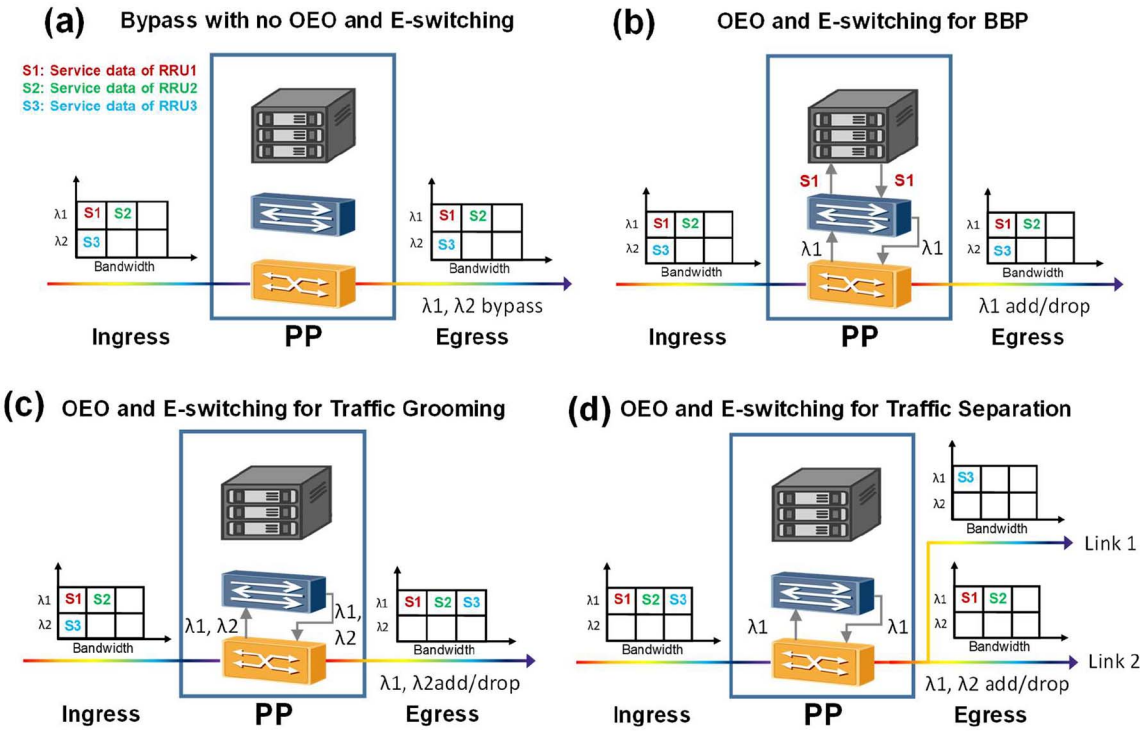

Fig. 4. The OEO conversion and E-switching operation for (a) bypass; (b) baseband processing; (c) traffic grooming; (d) traffic separation.

to aggregate traffics of both wavelengths onto a common one for bandwidth efficiency. 4) As shown in Fig. 4 (d), the OEO and E-switching is introduced for the traffic separation into different directions, in which *S*3 enters another ROAM port to another link. The latency value of switching in Fig. 4(c) and 4(d) are the same, but we distinguish them to represent two different switching situations.

(3) Latency for baseband processing. As reported in [28], BBU processing latency can be modelled as:

$$T(x, y, w)[\mu s] = c[x] + p[w] + u_r[x] + u_s(x, y) \quad (8)$$

where the triple (*x, y, w*) represents RB, MCS, and platform. The $c[x]$ and $u_r[x]$ is the latency of cell and user processing, respectively. The $u_s(x, y)$ is the specific user processing depending on RB and MCS. The $p[w]$ is the base offsets for running/operating the platform of BBU functions (i.e., virtualization environment). Thus, BBU processing latency is relevant with the BBP itself and virtualization environment. The latency of BBP itself depends on the CPU frequency that high-frequency CPU can achieve a shorter processing latency [25]. Therefore, we consider the BBP latency as a fixed value regardless of the split granularity. For example, if BBU is divided into five units, then the sum of the processing latency of five units is equal to one entire BBU. In contrast, we consider the virtualization environment latency in the model. The virtualization platform processing is assumed as T_V. The processing in one PP will introduce one platform latency. For example, if DU and CU are separated into two different PPs, then $2 \times T_V$ latency is introduced. So fine-grained split may introduce more platform latency than coarse-grained split.

(4) Latency for interface encapsulation. For fine-grained split, a reconfigurable and general interface (GI) for data encapsulation will be introduced. In 5G, the radio data is transmitted on a frame basis. So, we estimate the interface encapsulation latency referred to eCPRI [29]. The GI latency is defined as $T_{encap} = L_P/B_{Split}$, where L_P is the length (bit) of a frame, B_{split} is the data rate of split options. The GI encapsulation is applied to wireless data when finishing processing the last FU placed in this PP.

VI. MILP Formulations and Analysis

In this section, we propose a MILP model to determine the location of FUs to minimize resource consumption, latency, and network deployment cost. This MILP model can be adapted to all three split architectures.

A. MILP Model

In this model, we have set two objective functions (OF) to evaluate the placement performance. 1) The first OF is designed to minimize the number of used PPs, consumed bandwidth, and total transport latency. The latency includes the transmission latency for 5μs/km in fibers, latency of OEO and E-switching for 20μs [8], platform processing for 52μs [28]. 2) The second OF is designed to minimize the deployment cost, which comprises of PP cost and Path cost. The PP cost comprises the constant expenditure for housing/cooling/devices ($45000) [30] and variable expenditure described as dollar per GOPS for GPP ($1.59/GOPS) [31]. The path cost comprises the expenditure of fiber cable ($80/km) and fiber trenching/laying ($3000/km) [30].

1) Given:

Constant	Description
B, L, R	Set of RRUs, optical links, network nodes
K	Set of FU (e.g., $k = 2$ for FU1, $k = 1$ for RRU)
W	Set of wavelengths on each link
$F_{b,r}$	1 if RRU b is directly connected to PP *r*
E_r	1 if DC is placed at node *r*
$T^{i,j}$	The transmission latency of link *e(i, j)*, *i, j* $\in$ *R*
$M^{i,j}$	1 if PP *i* is directly connected to PP *j*
T_{soe}	Latency of OEO and E-switching operation
T_b	Latency requirement of services in RRU *b*
TC_b	Transmission Latency from RRU *b* to its directly connected PP
TE_k	GI encapsulation latency after FU *k* processed
$CK_{b,k}$	Computational demand of FU *k* of RRU *b*
$RB_{b,k}$	Bandwidth of RRU *b* after FU *k* processed
C_w	Capacity of wavelength *w*
C_r	Computational Capacity of PP *r*
Num	A large positive integer
sp	Number of FUs plus two (7 for FU, 3 for SBBU).

2) Variables:

Variable	Description
D_r	Binary, 1 if PP *r* is used
$H_{b,r}$	Binary, 1 if RRU *b* is processed in PP *r*
$Y_{b,r}$	Binary, 1 if RRU *b* is E-switched in PP *r*
$O_{b,k,r}$	Binary, 1 if FU *k* of RRU *b* is processed in PP *r*
$X_{b,k}^{i,j,w}$	Binary, 1 if RRU *b* is carried on wavelength *w* of link *e(i, j)* with the split state *k*
$Z_{b,k,r}$	Binary, 1 if FU *k* is the last unit processed in PP *r*
$P^{i,j}$	Binary, 1 if link *e(i, j)* is used.

3) Objective Function 1:

$$\begin{aligned} Minimize: a\sum_{r} D_r + b\sum_{\substack{i,j,b,\\k,w}} X_{b,k}^{i,j,w}\cdot RB_{b,k} \\ + c\left[\sum_{\substack{i,j,b,\\k,w}} X_{b,k}^{i,j,w}\cdot T^{i,j} + \sum_{r,b} Y_{b,r}\cdot T_{soe}\right. \\ \left. + \sum_{k,r} Z_{b,k,r}\cdot(T_v + 2\cdot TE_k)\right] \end{aligned} \tag{9}$$

The first part is to minimize the number of PPs, the second one is for bandwidth, and the third one is for latency. We set $a = 1/|R|$, $b = 1/(|L|\cdot|W|\cdot C_W)$, and $c = 1/(10^{\lceil log_{10}|R|\rceil}\cdot \sum_b T_b)$). The weight for latency is set far less than other two weights so that make the transport latency as the third objective because its upper bound is guaranteed in constraint (18).

4) Objective Function 2:

$$\begin{aligned} Minimize: (45000 + 1.59\cdot C_r)\cdot\sum_r D_r + (3080/5) \\ \times\left(\sum_{r,j>i} P^{i,j}\cdot T^{i,j}\right) \end{aligned} \tag{10}$$

The first part is to minimize the cost of PPs, while the second one is for the path cost. The value "3080/5" means that transmission latency is 5-fold than path length.

5) Constraints:

• *Routing:*

$$\begin{aligned} \sum_{i\neq r,k<\mathrm{sp},w} X_{b,k}^{i,r,w} - \sum_{j\neq r,k<sp,w} X_{b,k}^{r,j,w} \\ = \begin{cases} -1, & F_{b,r} = 1 \\ 1, & E_r = 1 \\ 0, & others \end{cases}, \forall b, r \end{aligned} \tag{11}$$

$$\sum_{k,w1\in W} X_{b,k}^{i,j,w1} + \sum_{l\in K,w2\in W} X_{b,l}^{j,i,w2} \leq 1,\ \forall i, j(i\neq j), b \tag{12}$$

The constraint (11) ensures that a lighpath is selected for each RRU from its directly connected PP to DC, while avoiding the loop formation through constraints (12).

• *Capacity:*

$$\sum_{b,k<sp} X_{b,k}^{i,j,w}\cdot RB_{b,k} \leq C_W,\ \forall i, j(i\neq j), w \tag{13}$$

$$\sum_{b,k<sp} O_{b,k,r}\cdot CK_{b,k} \leq C_r,\ \forall r \tag{14}$$

The constraint (13) ensures that data carried on each wavelength cannot exceed its capacity limitation, while computational capacity limitation is ensured in constraint (14).

• *Latency:*

$$\begin{aligned} \sum_{i,j,k<sp,w} X_{b,k}^{i,j,w}\cdot T^{i,j} + \sum_r Y_{b,r}\cdot T_{soe} + \sum_{k,r} Z_{b,k,r} \\ \times (T_v + 2\cdot TE_k) \\ + TC_b \leq T_b,\ \forall b \end{aligned} \tag{15}$$

The latency requirement of each RRU (from RRU to DC) is satisfied with constraint (15). The first part of Eqn. (15) is the transmission latency on fibers, the second one is for switching latency, the third one is for the platform processing and GI where "2" represents both encapsulation and decapsulation, the last is for the transmission from RRU to its directly connected PP.

• *FU Chain Placement:*

$$O_{b,k,r} = \begin{cases} F_{b,r}, & k = 1 \\ E_r, & k = sp \end{cases},\ \forall b, r \tag{16}$$

$$\sum_r O_{b,k,r} = 1,\ \forall b, k \tag{17}$$

$$\begin{aligned} O_{b,k,r} = \sum_{\substack{j,w\\ k\leq l\leq sp-1}} X_{b,l}^{r,j,w}\cdot M^{r,j},\ if\ F_{b,r} = 1, \\ \forall b, k, r \end{aligned} \tag{18}$$

$$\begin{aligned} O_{b,k,r} + 1 \geq \sum_{\substack{i,w,\\ 1\leq l\leq k-1}} X_{b,l}^{i,r,w}\cdot M^{i,r} \\ + \sum_{\substack{i,w,\\ k\leq n\leq sp-1}} X_{b,n}^{r,j,w}\cdot M^{r,j} \\ \geq 2\cdot O_{b,k,r},\ if\ F_{b,r}\neq 1,\ \forall b, k, r \end{aligned} \tag{19}$$

The constraint (16) describes the source and destination nodes for each RRU. The constraint (17) ensures that each FU is placed once in networks. The constraint (18) describes the situation that RRU b is processed in its directly connected PP, which means that split state (k) of RRU leaving this PP equals to the serial number of the last FU processed in this PP. The constraint (19) describes the situation that RRU b is processed in other PPs, which means that split state of RRU entering this PP equals to the serial number of the first FU processed in this PP, and split state of RRU leaving this PP equals to the serial number of the last processed FU.

• *OEO Conversion and E-Switching:*

$$\begin{aligned} Y_{b,r} \le & \sum_{b2\in B} \left(H_{b,r}\cdot H_{b2,r}\right) \\ & \times \left[\sum_{w,i\neq r}\left(\sum_k X_{b,k}^{i,r,w}\cdot\sum_k X_{b2,k}^{i,r,w}\right)\right. \\ & \left. + \sum_{w,j\neq r}\left(\sum_k X_{b,k}^{r,j,w}\cdot\sum_k X_{b2,k}^{r,j,w}\right)\right] \\ & \le Max_{num}\cdot Y_{b,r},\ \forall b,r \end{aligned} \tag{20}$$

Constraint (20) is to judge whether RRU b has experienced an OEO and E-switching in PP r. The $H_{b,r}\cdot H_{b2,r}$ is designed to judge whether RRU b or $b2$ or both are processed in PP r. The remaining part of Eqn. (20) is designed to judge whether RRU b and $b2$ are carried on the same wavelength to enter or leave PP r. If RRU b and $b2$ share the common wavelength and one of them is processed in PP r, then $Y_{b,r}=1$.

• *Others:*

$$H_{b,r} \le \sum_k O_{b,k,r} \le Num\cdot H_{b,r},\ \forall b, r\neq dc \tag{21}$$

$$D_r \le \sum_{b,1<k<sp} O_{b,k,r} \le Num\cdot D_r,\ \forall r \tag{22}$$

$$P^{i,j} \le \sum_{b,k,w}\left(X_{b,k}^{i,j,w}+X_{b,k}^{j,i,w}\right) \le Num\cdot P^{i,j},\ \forall i,j \tag{23}$$

$$\begin{aligned} 2\cdot Z_{b,k,r} &\le O_{b,k,r} - O_{b,k+1,r}+1 \\ &\le Z_{b,k,r}+1,\ \forall b, k\in(2,sp), r \end{aligned} \tag{24}$$

Constraint (21) is designed to judge whether RRU b is processed in PP r. Constraint (22) denotes whether PP r is used. Constraint (23) describes whether link $e(i,j)$ is used. Constraint (24) is designed to judge whether FU k is the last unit processed in PP r for RRU b. Eqn. (24) is relevant to the GI encapsulation.

Note that constraint (20) is a non-linear formula and cannot be solved via the optimization machinery. We have linearized it by using a series of linear constraints.

Linearization for OEO and E-Switching Constraints:

6) Variables:

Variable	Description
$I_{b1,b2}^r$	Binary, 1 if RRU $b1$ or $b2$ or both are processed in PP r
$Q_{b1,b2}^{i,j,w}$	Binary, 1 if RRU $b1$ and $b2$ share the same wavelength w on link $e(i,j)$
$Tin_{b1,b2}^{r,w}$	Binary, 1 if RRU $b1$ and $b2$ share the same wavelength to enter PP r, and $b1$ or $b2$ or both are processed in PP r
$Tout_{b1,b2}^{r,w}$	Binary, 1 if RRU $b1$ and $b2$ share the same wavelength to leave PP r, and $b1$ or $b2$ or both are processed in PP r
$G1_{b1,b2}^r$	Binary, auxiliary variable for *Gr b1,b2*
$G2_{b1,b2}^r$	Binary, auxiliary variable for *Gr b1,b2*
$G_{b1,b2}^r$	Binary, 1 if RRU $b1$ and $b2$ enter PP r on the same wavelength but leave with different wavelengths, or enter PP r on different wavelengths but leave with the same one.

7) Constraints:

• Situation in Fig. 4 (b)

$$I_{b1,b2}^r \le H_{b1,r}+H_{b2,r} \le 2\cdot I_{b1,b2}^r,\quad \forall b1,b2\in B, r\in R \tag{25}$$

$$2\cdot Q_{b1,b2}^{i,j,w} \le \sum_k X_{b1,k}^{i,j,w}+\sum_l X_{b2,l}^{i,j,w} \le Q_{b1,b2}^{i,j,w}+1,\quad \forall b1,b2,i,j,w \tag{26}$$

$$2\cdot Tin_{b1,b2}^{r,w} \le I_{b1,b2}^r+\sum_{i\neq r} Q_{b1,b2}^{i,r,w} \le Tin_{b1,b2}^{r,w}+1,\quad \forall b1,b2,r,w \tag{27}$$

$$2\cdot Tout_{b1,b2}^{r,w} \le I_{b1,b2}^r+\sum_{i\neq r} Q_{b1,b2}^{r,j,w} \le Tout_{b1,b2}^{r,w}+1,\quad \forall b1,b2,r,w \tag{28}$$

Constraint (20) can be divided into two parts for linearization, i.e., FU processing (Constraints (25)–(28)), and E-switching (Constraints (29) – (31)). Constraint (25) denotes whether RRU $b1$ or $b2$ or both are processed in PP r. Constraint (26) describes whether RRU $b1$ and $b2$ share the common wavelength on link $e(i,j)$. Constraints (27) and (28) ensure that all RRUs carried on wavelength w (ingress or egress) will experience an OEO and switching when one of these RRUs are processed in PP r.

• Situations in Fig. 5 (c)∼(d)

$$\begin{aligned} G1_{b1,b2}^r &\ge -3+2\cdot\sum_{i,w} Q_{b1,b2}^{i,r,w}+2\cdot\sum_{j,w} Q_{b1,b2}^{r,j,w} \\ &\ge 4\cdot G1_{b1,b2}^r-3,\ \forall b1,b2,r \end{aligned} \tag{29}$$

$$G2_{b1,b2}^r \ge 1-2\cdot\sum_{i,w} Q_{b1,b2}^{i,r,w}+\sum_{j,w} Q_{b1,b2}^{r,j,w} \ge 4\cdot G2_{b1,b2}^r-3,\quad \forall b1,b2,r \tag{30}$$

$$G_{b1,b2}^r = 1-\left(G1_{b1,b2}^r+G2_{b1,b2}^r\right),\ \forall b1,b2,r \tag{31}$$

Constraints (29)–(31) are designed to denote whether traffic grooming or separation have occurred, which can be known through a comparison of contained RRUs on one wavelength when entering and leaving PP r. If contained RRUs change (e.g., some RRU joins or leaves), grooming or separation has happened and introduced an OEO and E-switching operation. For example, if *RRU 1, 2, 3* and *RRU 2, 3, 4* are respectively carried on wavelength w when entering and leaving PP r, then contained RRUs have changed because *RRU 1* has left w with *RRU 4* joining in them. Constraint (29) and (30) are auxiliary constraints to bridge the ($Q_{b1,b2}^{i,r,w}$, $Q_{b1,b2}^{r,j,w}$) pair and $G_{b1,b2}^r$.

• Judging whether OEO and E-switching occurs

$$Num \cdot Y_{b,r} \geq \sum_{b2,w1} Tin^{r,w1}_{b,b2} + \sum_{b3,w2} Tout^{r,w2}_{b,b3} + \sum_{b4} G^{r}_{b,b4} \geq Y_{b,r}, \forall b, r \quad (32)$$

$$Y_{b,r} \geq \sum_{i,k<sp} X^{i,r,w}_{b,k} - \sum_{j,l<sp} X^{r,j,w}_{b,l} \geq -1 \cdot Y_{b,r},\ \forall E_r \neq 1, b, w \quad (33)$$

Constraint (32) describes whether there is an OEO and E-switching operation for RRU b in PP r. Constraint (33) ensures that bypassing data shouldn't change its wavelength.

B. Analysis of Relationship Between BBP Centralization and Split Granularity

As mentioned above, future RAN will significantly benefit from the BBP centralization, which dominates the network expenditure and power consumption. But how split granularity influences the centralization gain is still a problem that deserves our further discussion. In this section, we will present a theoretical analysis on the centralization performance under different split granularity.

Here, we consider a limit case of a fully flexible functional split (FFS) where BBU can be divided into units of arbitrary size. Although this case is somehow unreasonable from the perspective of actual baseband processing, it really reflects the ultimate advantage of the fine-grained split. The RRU-SBBU is provided as a benchmark because of its non-split for remaining functions.

First, we evaluate the centralization gain of FFS and SBBU by calculating the respective number of used PPs (NUP) in Eqn. (34) and Eqn. (35). For ease of analysis, bandwidth on each link is assumed to be sufficient. And the non-sufficient case will show a similar trend subject to its bandwidth capacity. The parameters *N, M, C* represent the number of RRUs, computational demands per RRU, and computational capacity per PP. In Eqn. (34), the symbol $\lceil\cdot\rceil$ represents the operation for taking the smallest integer larger than •, while $\lceil\cdot\rceil$ is the opposite.

$$NUP_{FFS} = \left\lceil \frac{NC}{M} \right\rceil,\ NUP_{SBBU} = \left\lceil \frac{N}{\lfloor M/C \rfloor} \right\rceil \quad (34)\ \&(35)$$

To evaluate the NUP, we define $\mathrm{M} = \rho \cdot \mathrm{C}$ ($\rho \in \mathrm{R}+$), where ρ is defined as the ratio of PP capacity to computational demand per RRU, to convert above formulas as follows:

$$NUP_{FFS} = \left\lceil \frac{N}{\rho} \right\rceil,\ NUP_{SBBU} = \left\lceil \frac{N}{\lfloor \rho \rfloor} \right\rceil \quad (36)\ \&\ (37)$$

Given the computational complexity of $\lfloor\cdot\rfloor$ operation, we then define $\rho = a \cdot N + b$ ($a \in Z,\ b \in [0,\ N]$) and $\lfloor\rho\rfloor = a \cdot N + b - s$($\mathbf{s}$ is the fractional part of $\boldsymbol{\rho}, \mathbf{s} \in [\mathbf{0},\ \mathbf{1}]$) for the following comparison.

$$\frac{NUP_{FFS}}{NUP_{SBBU}} = \frac{\lceil 1/(a+b)/N \rceil}{\lceil 1/(a+(b-s))/N \rceil} \quad (38)$$

Concluded from the above analysis, we can obtain the following result in Eqn. (39), and then convert Eqn. (38) to Eqn. (40).

$$\left\lceil \frac{1}{a+b/N} \right\rceil \leq \left\lceil \frac{1}{a+(b-s)/N} \right\rceil \leq \left\lceil \frac{1}{a+(b-1)/N} \right\rceil \quad (39)$$

$$1 \geq \frac{NUP_{FFS}}{NUP_{SBBU}} \geq \frac{\lceil 1/(a+b/N) \rceil}{\lceil 1/(a+(b-1)/N \rceil} \quad (40)$$

We assume that number of RRUs (N) is already known in our discussion and the right side of inequality is relevant to a and b, which increases with the increment of ρ. The piecewise function is introduced to elaborate the change of NAP_{FFS}/NAP_{SBBU} along with ρ. **1)** $\boldsymbol{\rho}$<**N**. Under this circumstance, we can calculate that parameter $\boldsymbol{a}$ equals zero. The right side of inequality can be converted to $\lceil \frac{N}{b} \rceil / \lceil \frac{N}{b-1} \rceil$ in Eqn. (41). The difference between $\lceil \frac{N}{b} \rceil$ and $\lceil \frac{N}{b-1} \rceil$ becomes smaller with the increment of b (i.e., ρ). **2)** $\boldsymbol{\rho} \geq$ **N**. Under this circumstance, parameter $\boldsymbol{a}$ is larger than "1". We can calculate that $\lceil \frac{1}{a+b/N} \rceil$ and $\lceil \frac{1}{a+(b-1)/N} \rceil$ both equal "1" in Eqn. (42), which means that the number of used PPs in FFS and SBBU are the same.

$$\rho < N : 1 \geq \frac{NAP_{FFS}}{NAP_{SBBU}} \geq \frac{\lceil N/b \rceil}{\lceil N/(b-1) \rceil} \quad (41)$$

$$\rho \geq N : \frac{NAP_{FFS}}{NAP_{SBBU}} = 1 \quad (42)$$

Therefore, we can obtain that advantage on centralization gain from fine-grained split is significantly obvious in low-ρ scenario but decreases generally with the increment of ρ. Thus, there are two factors influencing the centralization gain, i.e., split granularity and ρ. Fig. 5 presents the relationship between the granularity of split, ρ, and number of used PPs, which has also been confirmed in the simulation.

VII. Performance Evaluation

A. Simulation Setup

In the simulation, each RRU contains four antennas, two MIMO layers, a 100MHz wireless spectrum (i.e., 500 RBs), and MCS 23. Thus, calculated from formulations in Section IV, the total computational demand of one RRU is about 1800 GOPS. The latency requirement of each RRU is set as 500 μs (not contain the processing of BBP itself). The MILP formulations are programmed on IBM CPLEX 12.7 software, which runs on a high-performance server of 32G RAM.

We have designed two simulation scenarios in the following part: 1) small-scale network with 1∼8 RRUs to compare different ρ scenarios, and 2) then extend the comparison to a large-scale network with 10∼40 and 10∼80 RRUs.

B. Simulation Scenario I

As shown in Fig. 6, we consider a network topology of 8 PP nodes [7], 1 DC node, and 16 optical links. Each optical link ranges in [5], [30] km, while the distance between RRU and its directly connected PP ranges in [0, 3] km. The *PP 1, 2* are assumed to be directly connected with RRUs (up to 4 RRUs per PP). The data center is located at *Node 9*. Each optical link runs at 50Gbps capacity. We have studied on the placement performance under different ρ ($\rho = 1.5 \rightarrow 2700$ GOPS,

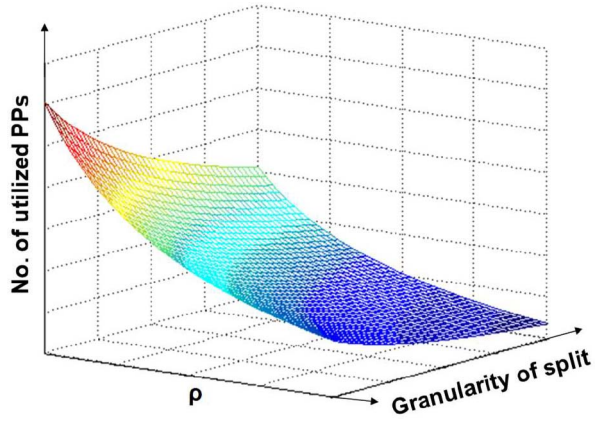

Fig. 5. Diagram of relationship between granularity of split, ρ, and No. of PPs.

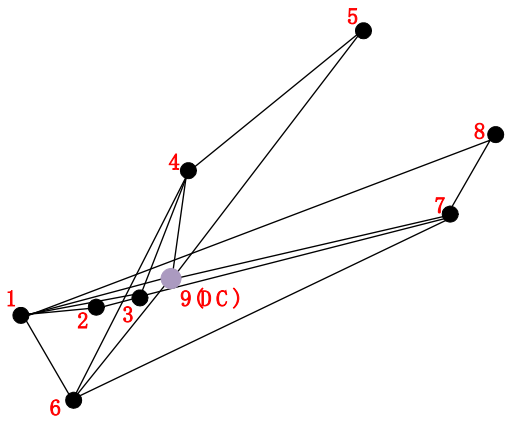

Fig. 6. Simulation topology for small-scale network [7].

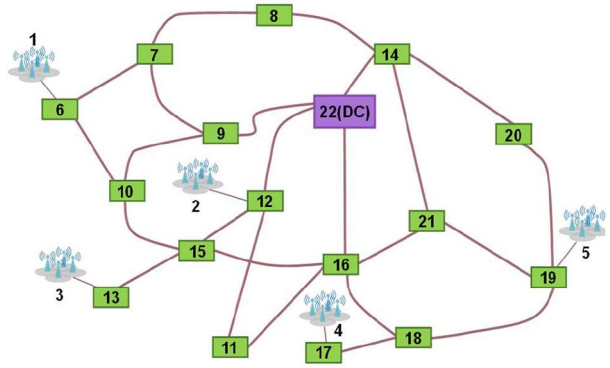

Fig. 7. Simulation topology for large-scale network.

$\rho = 1.75 \rightarrow 3150$ GOPS, $\rho = 2.5 \rightarrow 4500$ GOPS). For each ρ, we fix the computational and link capacity in the network and increase the number of RRUs.

We compare the three split architectures in six aspects in the small-scale network.

1) Number of Used PPs vs. Number of RRUs: The PP selection is an important property for network operators to consider. As shown in Fig. 8 (a), 9(a), and 10(a), the number of PPs increases with RRUs in all three architectures. Because the increment of RRU requires more computational resources and then activates more PPs. RRU-FU obtains the optimal performance because it divides BBU into multiple fine-grained units to promote the centralization gain (seen as a bin-packing problem). RRU-DU-CU achieves a higher centralization gain compared with RRU-SBBU due to its two splits. We can also find that gaps between three architectures become smaller with the increment of ρ, which is consistent with our analysis in Part B of Section VI.

2) Bandwidth vs. Number of RRUs: Optical bandwidth is a scarce resource that should be evaluated. We calculate the summed bandwidth consumption on all links for three architectures. As shown in Fig. 8(a), 9 (a), and 10 (a), FU saves 72% and 61% bandwidth than SBBU and DU-CU on average, respectively. That's because bandwidth decreases with baseband processing layer-by-layer, e.g., *split 7-2* requires 16-fold bandwidth than *split PDCP-RLC*. Fine-grained split can provide more choices in wavelength and PP allocation so that high-bandwidth functions can be processed in nearby PPs of each RRU (i.e., avoid the multiple-hop and long-reach transmission). Through the simulation, it can be confirmed that the fine-grained split can promote optical bandwidth saving.

3) Latency vs. Number of RRUs: The transport latency may influence the service experience that should also be optimized. We can observe in Fig. 8(b), 9(b), 10(b) that latency increases with the number of RRUs because of the more transmitted data. We can also observe that RRU-SBBU needs the minimum latency followed by RRU-DU-CU and RRU-FU. That's because functional split introduces more switching and virtualization platform processing latency. We can also observe that FU needs shorter latency than DU-CU at low network load. For example, FU requires shorter latency when there are less than 5 RRUs in Fig. 9 (b). This conclusion is also the same in Fig. 10(b) when there are less than 8 RRUs (only 4 PPs are activated for 8 RRUs, so it is still at a low network load). The reason for this phenomenon is that FU can provide more choices in routing at low network load because of its fewest bandwidth consumption.

4) Number of OEO and E-Switching vs. Number of RRUs: As showed in Fig. 8(c), 9(c), 10(c), we can observe that RRU-SBBU requires fewest OEO and E-switching operations. That's because RRU-FU and RRU-DU-CU introduce the frequent baseband processing that results in extra OEO and E-switching as described in Fig. 4(b).

5) Number of Splits vs. Number of RRUs: As showed in Fig. 8(d) and 9(d), RRU-FU requires the most splits to achieve the highest centralization gain. In Fig. 10(d), RRU-DU-CU requires slightly more splits compared with RRU-FU at low network load. That's because all DU-CU pairs are separated into two PPs, but not all FU sets are separated which derives from its more available split options.

6) Network Deployment Cost vs. Number of RRUs: Operators show great attention on network cost that deserves our further discussion. In Fig. 11, we can observe that RRU-FU can achieve the best cost budget because of its centralization gain. RRU-FU saves 23.6% and 14.7% expenditure than RRU-SBBU and RRU-DU-CU on average, respectively. Moreover, the advantage of bandwidth consumption also contributes to path economization. That's because the path

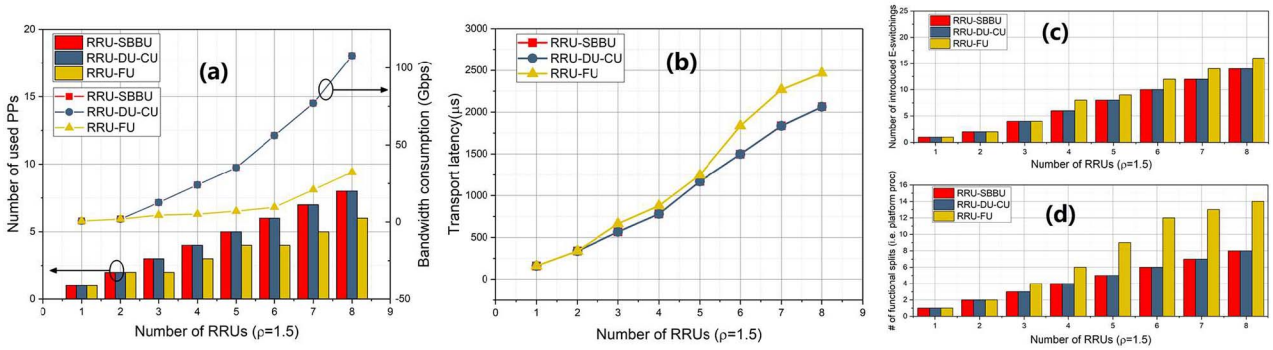

Fig. 8. (a) Number of PPs (bandwidth) vs. RRUs, (b) Latency vs. RRUs, (c) Number of OEO and E-switching vs. RRUs, (d) Number of functional splits (i.e., platform processing) vs. RRUs for $\rho = 1.5$ in small-scale network.

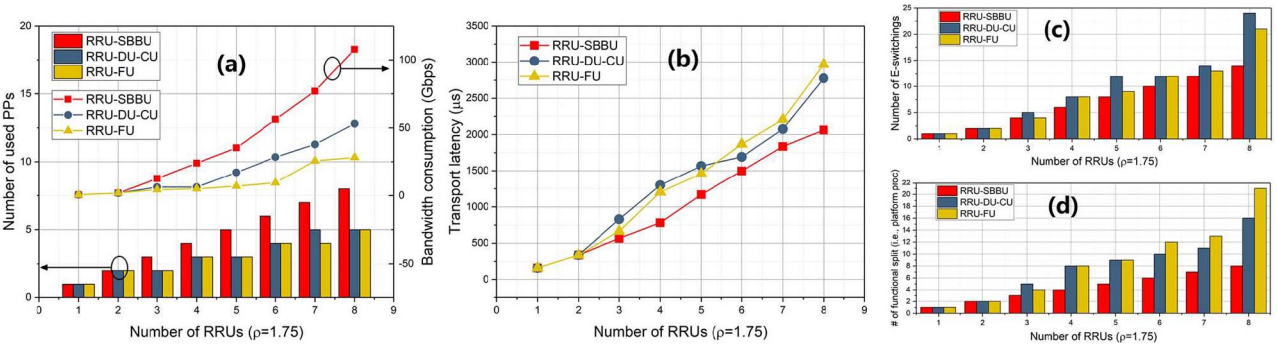

Fig. 9. (a) Number of PPs (bandwidth) vs. RRUs, (b) Latency vs. RRUs, (c) Number of OEO and E-switching vs. RRUs, (d) Number of functional splits (i.e., platform processing) vs. RRUs for $\rho = 1.75$ in small-scale network.

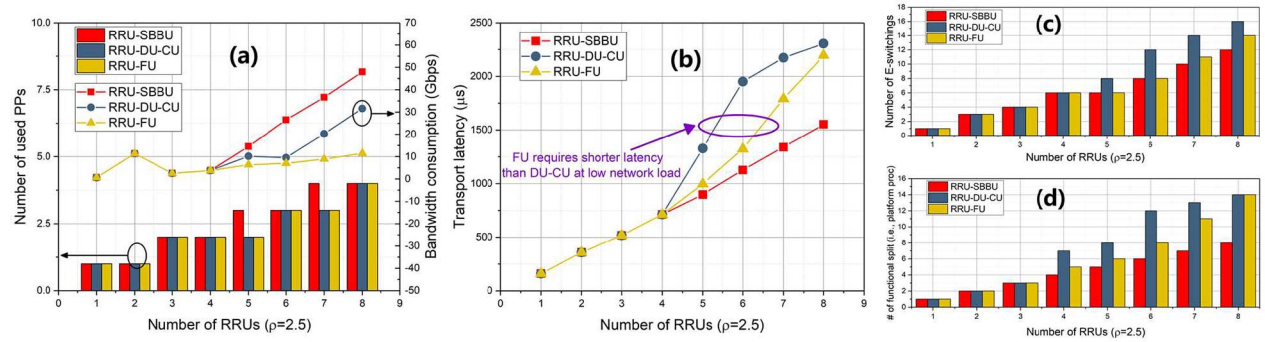

Fig. 10. (a) Number of PPs (bandwidth) vs. RRUs, (b) Latency vs. RRUs, (c) Number of OEO and E-switching vs. RRUs, (d) Number of functional splits (i.e., platform processing) vs. RRUs for $\rho = 2.5$ in small-scale network.

selection can be more flexible, and each activated link can carry more services.

C. Simulation Scenario II

As shown in Fig. 7, we consider a large-scale network scenario, which contains 16 PP nodes, 5 RRU nodes, and 28 optical links. The DC is located at *Node 22*. Each optical link ranges in [10], [30] km which are assumed to provide 200 Gbps capacity. Also, we compare three architectures under two ρ scenarios ($\rho = 3.75 \rightarrow 6750$ GOPS, $\rho = 6.75 \rightarrow 12150$ GOPS) on six aspects in the large-scale network. There are 5$\sim$40 RRUs in $\rho = 3.75$ scenario and 10$\sim$80 RRUs in $\rho = 6.75$ scenario.

1) Number of Used PPs vs. Number of RRUs: As shown in Fig. 12(a) and 13(a), RRU-FU shows the highest centralization gain compared with the other two architectures. The result is consistent with that in the small-scale network.

2) Bandwidth vs. Number of RRUs: As shown in Fig. 12(a) and 13(a), RRU-FU consumes the minimal bandwidth in most cases. In Fig. 12(a), we can observe that RRU-FU saves 44% and 34% bandwidth than RRU-SBBU and RRU-DU-CU on average, respectively. In Fig. 13(a), RRU-FU saves 52% and 43% bandwidth than RRU-SBBU and RRU-DU-CU on average, respectively. The reason is consistent with that in the small-scale network. However, at 10-RRUs and 15-RRUs simulation node in Fig. 12(a), RRU-FU requires more bandwidth because of its PP saving at low network load. The PP saving has resulted in the long-reach and multi-hop transmission for some RRUs to remote PPs that consumes more bandwidth.

3) Latency vs. Number of RRUs: As shown in Fig. 12(b) and 13(b), RRU-SBBU achieves the shortest latency followed

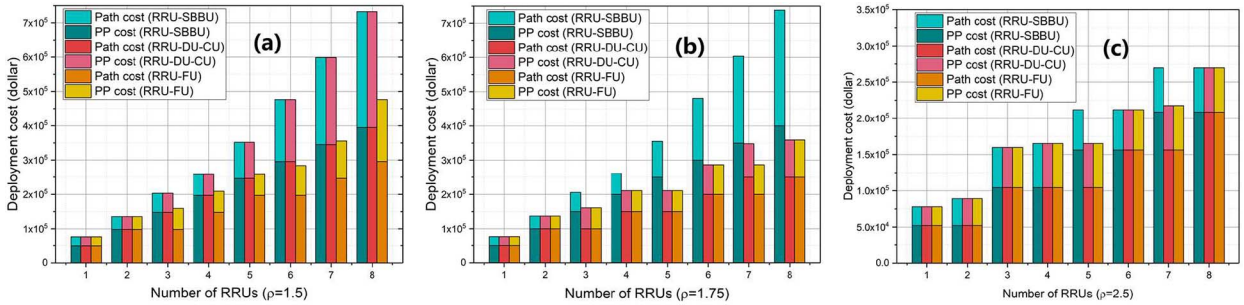

Fig. 11. Network deployment cost vs. RRUs at (a) $\rho = 1.5$, (b) $\rho = 1.75$, (c) $\rho = 2.5$ in small-scale network.

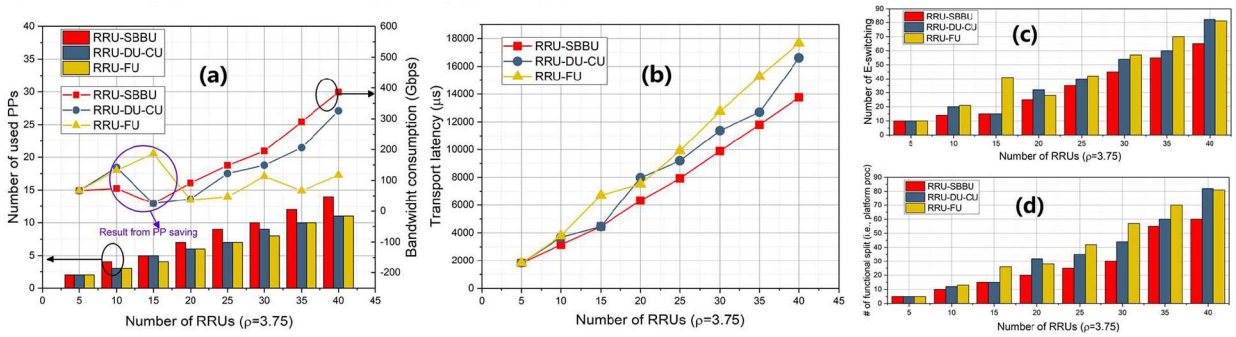

Fig. 12. (a) Number of PPs (bandwidth) vs. RRUs, (b) Latency vs. RRUs, (c) Number of OEO and E-switching vs. RRUs, (d) Number of functional splits (i.e., platform processing) vs. RRUs for $\rho = 3.75$ in large-scale network.

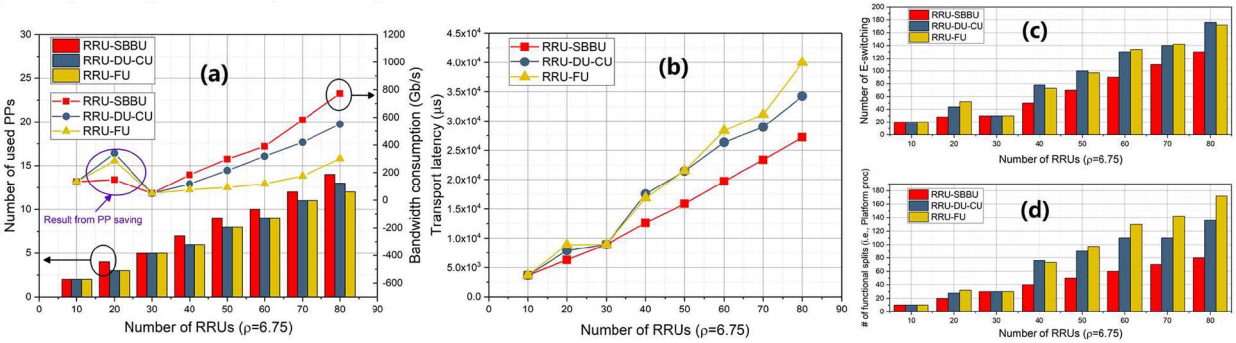

Fig. 13. (a) Number of PPs (bandwidth) vs. RRUs, (b) Latency vs. RRUs, (c) Number of OEO and E-switching vs. RRUs, (d) Number of functional splits (i.e., platform processing) vs. RRUs for $\rho = 6.75$ in large-scale network.

by RRU-DU-CU and RRU-FU. In Fig. 12(b), we can observe that RRU-FU requires 21.6% and 10.1% more latency than RRU-SBBU and RRU-DU-CU, respectively. In Fig. 13(b), RRU-FU needs 26.1% and 16.3% more latency than RRU-SBBU and RRU-DU-CU, respectively. That's because FU benefits from the functional split but requires more distributed baseband processing during its whole BBP chain, and then introduces extra OEO and E-switching and platform processing.

4) Number of OEO and E-Switching and Splits vs. Number of RRUs: Fig. 12(c) and 13(c) show the number of OEO and E-switching in three split architectures. From numerical results in Fig. 12(c), we can observe that RRU-FU and RRU-DU-CU introduce 55.6% and 36.9% more operations than RRU-SBBU on average. In Fig. 13(c), we can observe that RRU-FU and RRU-DU-CU introduce 36.3% and 35.9% more operations than RRU-SBBU. That's because more OEO and E-switching is resulted from the distributed baseband processing.

Moreover, we can observe in Fig. 12(d) that 78.9% and 38.9% more platform processing are introduced by FU and DU-CU on average, respectively. In Fig. 13(d), 90.5% and 63.8% more platform processing are introduced by FU and DU-CU. That's because PP centralization benefits from the distributed baseband processing, where more virtualization platform processing is also introduced.

5) Network Deployment Cost vs. Number of RRUs: As shown in Fig. 14-15, RRU-FU achieves the best cost budget followed by RRU-DU-CU and RRU-SBBU. From the numerical results, RRU-FU can economize the expenditure of 21.5% (10.1% in Fig. 15) than RRU-SBBU and 8.2% (5.3% in Fig. 15) than RRU-DU-CU on average.

VIII. Conclusion

We provide a detailed discussion on RAN deployment under the fine-grained split architecture with our proposed MILP

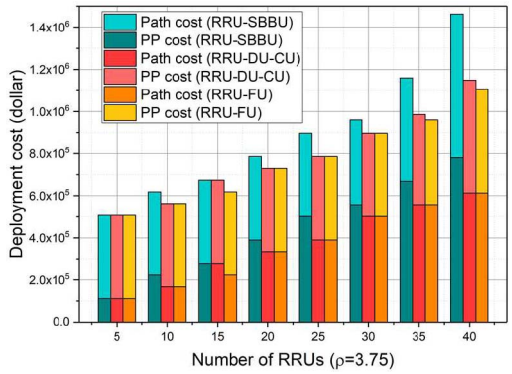

Fig. 14. Deployment cost vs. RRUs at $\rho = 3.75$ in large-scale network.

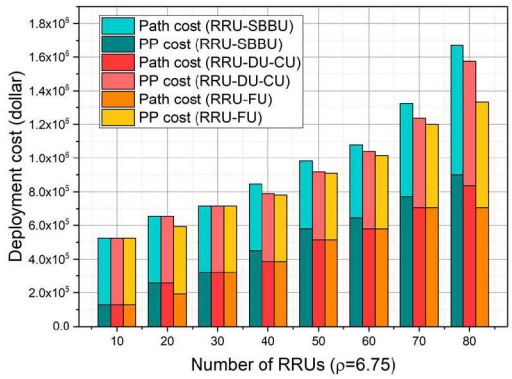

Fig. 15. Deployment cost vs. RRUs at $\rho = 6.75$ in large-scale network.

model. We compare the RRU-FU architecture with RRU-SBBU and RRU-DU-CU in terms of the resource efficiency, deployment cost, and latency. The simulation results show that RRU-FU can achieve the best performance on PP centralization, bandwidth saving (over 40% and 30% reduction than SBBU and DU-CU), and network deployment economization ($\sim$18.4% and 9.4% economization than SBBU and DU-CU). However, RRU-FU may require more latency than coarse-grained split (over 20% and 10% more latency than SBBU and DU-CU in the large-scale network) because of introducing more OEO and E-switching operations and virtualization platform processing. With the results and analysis of the simulation, we conclude some pros and cons of fine-grained split architecture for summarization.

- *Pros:* 1) The fined-grained split (FGS) can balance the trade-off between the BBP centralization gain and optical bandwidth saving. 2) FGS can contribute to the deployment cost economization because of using few PPs and fiber cables. 3) FGS can benefit the isolation between network slices by using a finer virtual network function. For example, if the MAC layer of *slice1* should be isolated with other slices, then only the MAC layer will be dependently placed, and other functions can still share the computational resource with other slices.
- *Cons:* 1) A general and re-configurable interface should be designed between any adjacent FUs, which may increase the transport latency because of the frequent encapsulation and decapsulation. 2) More latency for virtualization platform processing may be introduced that a lightweight virtualization technology should be designed to support the fine-grained split. 3) The complicated network element control and management will be introduced to orchestrate and monitor these independent units.

REFERENCES

[1] *Technical Specification Group Services and System Aspects; Service Requirements for the 5G System, V17.1.0, Rel. 17*, 3GPP Standard TS 22.261, Dec. 2019.

[2] I. A. Alimi, A. L. Teixeira, and P. P. Monteiro, "Toward an efficient C-RAN optical fronthaul for the future networks: A tutorial on technologies, requirements, challenges, and solutions," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 1, pp. 708–769, 1st Quart., 2017.

[3] Y. Ji, J. Zhang, X. Wang, and H. Yu, "Towards converged, collaborative and co-automatic (3C) optical networks," *Sci. China Inf. Sci.*, vol. 61, no. 12, Nov. 2018, Art. no. 121301.

[4] Y. Ji, J. Zhang, Y. Xiao, and Z. Liu, "5G flexible optical transport networks with large-capacity, low-latency and high-efficiency," *China Commun.*, vol. 16, no. 5, pp. 19–32, May 2019.

[5] *Technical Specification Group Radio Access Network; NG-RAN; Architecture Description, V16.0.0, Rel. 16*, 3GPP Standard TS 38.401, Dec. 2019.

[6] "Study on new radio access technology: Radio access architecture and interfaces, V14.0.0, rel. 14," 3GPP, Sophia Antipolis, France, Rep. 38.801, Mar. 2017.

[7] F. Musumeci, C. Bellanzon, N. Carapellese, M. Tornatore, A. Pattavina, and S. Gosselin, "Optimal BBU placement for 5G C-RAN deployment over WDM aggregation networks," *J. Lightw. Technol.*, vol. 34, no. 8, pp. 1963–1970, Apr. 15, 2016.

[8] Y. Li *et al.*, "Joint optimization of BBU pool allocation and selection for C-RAN networks," in *Opt. Fiber Commun. Conf. Expo. Opt. Soc. Amer. Tech. Dig. (OSA OFC)*, San Diego, CA, USA, 2018, pp. 1–3.

[9] F. Musumeci, G. Belgiovine, and M. Tornatore, "Dynamic placement of baseband processing in 5G WDM-based aggregation networks," in *Opt. Fiber Commun. Conf. Expo. Opt. Soc. Amer. Tech. Dig. (OSA OFC)*, San Francisco, CA, USA, 2017, p. M2G-4.

[10] O. Arouk, T. Turletti, N. Nikaein, and K. Obraczka, "Cost optimization of cloud-RAN planning and provisioning for 5G networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kansas City, MO, USA, 2018, pp. 1–6.

[11] C.-Y. Chang *et al.*, "Slice orchestration for multi-service disaggregated ultra-dense RANs," *IEEE Commun. Mag.*, vol. 56, no. 8, pp. 70–77, Aug. 2018.

[12] Y. Xiao, J. Zhang, Z. Gao, and Y. Ji, "Service-oriented DU-CU placement using reinforcement learning in 5G/B5G converged wireless-optical networks," in *Opt. Fiber Commun. Conf. Expo. Opt. Soc. Amer. Tech. Dig. (OSA OFC)*, San Diego, CA, USA, 2020, p. T4D-5.

[13] D. Harutyunyan and R. Riggio, "Flex5G: Flexible functional split in 5G networks," *IEEE Trans. Netw. Serv. Manag.*, vol. 15, no. 3, pp. 961–975, Sep. 2018.

[14] X. Wang, L. Wang, S. E. Elayoubi, A. Conte, B. Mukherjee, and C. Cavdar, "Centralize or distribute? A techno-economic study to design a low-cost cloud radio access network," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Washington, DC, USA, 2017, pp. 1–7.

[15] A. Tzanakaki, M. P. Anastasopoulos, and D. Simeonidou, "Optical networking: An important enabler for 5G," in *Proc. Eur. Conf. Opt. Commun. (ECOC)*, Gothenburg, Sweden, 2017, pp. 1–3.

[16] J. Zhang, Y. Ji, S. Jia, H. Li, X. Yu, and X. Wang, "Reconfigurable optical mobile fronthaul networks for coordinated multipoint transmission and reception in 5G," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 9, no. 6, pp. 489–497, Jun. 2017.

[17] J. Zhang, Y. Ji, H. Yu, X. Huang, and H. Li, "Experimental demonstration of fronthaul flexibility for enhanced CoMP service in 5G radio and optical access networks," *Opt. Exp.*, vol. 25, no. 18, pp. 21247–21258, Sep. 2017.

[18] X. Sun and N. Ansari, "Latency aware workload offloading in the cloudlet network," *IEEE Commun. Lett.*, vol. 21, no. 7, pp. 1481–1484, Jul. 2017.

[19] A. Garcia-Saavedra, J. X. Salvat, X. Li, and X. Costa-Perez, "WizHaul: On the centralization degree of cloud RAN next generation fronthaul," *IEEE Trans. Mobile Comput.*, vol. 17, no. 10, pp. 2452–2466, Oct. 2018.

[20] *Technical Specification Group Radio Access Network; NR; Medium Access Control (MAC) Protocol Specification, V15.8.0, Rel. 15*, 3GPP Standard TS 38.321, Dec. 2019.
[21] *Technical Specification Group Radio Access Network; NR; Radio Link Control (RLC) Protocol Specification V15.5.0, Rel. 15*, 3GPP Standard TS 38.322, Mar. 2019.
[22] *Technical Specification Group Radio Access Network; NR; Packet Data Convergence Protocol (PDCP) Specification V15.6.0, Rel. 15*, 3GPP Standard TS 38.323, Jun. 2019.
[23] M. Shehata, A. Elbanna, F. Musumeci, and M. Tornatore, "Multiplexing gain and processing savings of 5G radio-access-network functional splits," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 4, pp. 982–991, Dec. 2018.
[24] *Technical Specification Group Radio Access Network; NR; Physical Layer Procedures for Data, V16.0.0, Release 16*, 3GPP Standard TS 38.214, Dec. 2019.
[25] S. Khatibi, K. Shah, and M. Roshdi, "Modelling of computational resources for 5G RAN," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Ljubljiana, Slovenia, 2018, pp. 1–5.
[26] *Small Cell Virtualization Functional Splits and Use Cases*, document SCF159, Small Cell Forum Release, London, U.K., 2016.
[27] L. M. Larsen, A. Checko, and H. L. Christiansen, "A survey of the functional splits proposed for 5G mobile crosshaul networks," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 146–172, 1st Quart., 2018.
[28] N. Nikaein, "Processing radio access network functions in the cloud: Critical issues and modeling," in *Proc. 6th Int. Workshop Mobile Cloud Comput. Serv.*, Paris, France, 2015, pp. 36–43.
[29] M. Waqar and A. Kim, "Performance improvement of Ethernet-based fronthaul bridged networks in 5G cloud radio access networks," *Appl. Sci.*, vol. 9, no. 14, p. 2823, Jul. 2019.
[30] X. Wang, Y. Ji, J. Zhang, L. Bai, and M. Zhang, "Joint optimization of latency and deployment cost over TDM-PON based MEC-enabled cloud radio access networks," *IEEE Access*, vol. 8, pp. 681–696, 2019.
[31] S. McIntosh-Smith, J. Price, T. Deakin, and A. Poenaru, "A performance analysis of the first generation of HPC-optimized arm processors," *Concurrency Comput. Pract. Exp.*, vol. 31, no. 16, Aug. 2019, Art. no. e5110.

Yuming Xiao (Graduate Student Member, IEEE) received the bachelor's degree in information engineering and the master's degree in electronic and communication engineering from the Beijing University of Posts and Telecommunications, in 2014 and 2017, respectively, where he is currently pursuing the Ph.D. degree. His research interests mainly cover the networking and resource optimization in mobile fronthaul, midhaul, and backhaul networks.

Jiawei Zhang (Member, IEEE) received the Ph.D. degree from the State Key Laboratory of Information Photonics and Optical Communications, Beijing University of Posts and Telecommunications, Beijing, China, where he is currently an Associate Professor. He was a joint-supervised Ph.D. student with the University of California at Davis, Davis, CA, USA. His current research interests include 5G RAN transport networks, network function virtualization, software defined radio, and optical access networks, with an emphasis on the various fronthaul-/midhaul-/backhaul-based optical communication technologies.

Yuefeng Ji (Senior Member, IEEE) received the Ph.D. degree from the Beijing University of Posts and Telecommunications, where he is currently a Professor and the Deputy Director of the State Key Lab of Information Photonics and Optical Communications. His research interests are primarily in the area of broadband communication networks and optical communications, with emphasis on key theory, realization of technology, and applications. He is a Fellow of the China Institute of Communications, China Institute of Electronic, and IET.