IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Digital Object Identifier

FlexDRAN: Flexible Centralization in Disaggregated Radio Access Networks

Chia-Yu Chang[1], Navid Nikaein[2], Thrasyvoulos Spyropoulos[2], Koen De Schepper[1]
[1]Nokia Bell Labs, 2018 Antwerp, Belgium
[2]EURECOM, 06904 Sophia-Antipolis, France

Corresponding author: Chia-Yu Chang (e-mail:chia-yu.chang@nokia-bell-labs.com)

This work is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement 5Growth (No. 856709), Affordable5G (No. 957317), DAEMON (No. 101017109), and 6GBrain (No. 101017226).

ABSTRACT Radio Access Network (RAN) disaggregation allows operators to mix-and-match multivendor components and bring RAN services from one end to the other. Despite this goal, issues of resource misuse or performance undershoot may arise because of inflexible RAN function deployment and uncoordinated decision-making across different network segments. To address these issues, this paper considers full flexibility in the synthesis of end-to-end RAN services from a set of disaggregated and uncoordinated components. In particular, five design factors are jointly considered to maximize the overall network spectral efficiency: (1) User association, (2) Remote radio unit clustering, (3) RAN functional split, (4) Fronthaul network routing, and (5) Baseband unit placement. To efficiently deal with the formulated problem, we propose a two-level turbo-based solution and compare its performance with several related works. The simulation results show that our proposed solution can not only achieve a 1.33-times spectral efficiency gain compared with state-of-the-art methods, but also provides 1.27 and 1.74 multiplexing benefits for computing and networking resources, respectively.

INDEX TERMS Radio access networks, 5G mobile communication, Algorithm design and analysis, Disaggregated network, Functional split

I. INTRODUCTION

RADIO Access Network (RAN) is traditionally planned in a distributed and decentralized manner, and thus it has been referred to as Distributed RAN (D-RAN) since the 3G/4G era. However, some disadvantages of D-RAN, such as being incapable of being promptly reconfigured for new use cases and being flexibly coordinated among a number of Base Station (BS) processing, hinder its direct application in the Fifth-Generation (5G) system. Even worse, its one-size-fits-all solution will significantly increase the control and management overhead when serving multiple services under different sharing models between operators/vendors. To this end, an evolution toward Cloud/Centralized RAN (C-RAN) stands out as a promising solution, as mentioned in [1], [2].

Originally, the C-RAN prototype realizes both efficient network management and coordinated processing by replacing a monolithic BS with passive radio elements at cell sites, called Remote Radio Heads (RRHs), and a centralized pool of BaseBand Units (BBUs), where the baseband and protocol processing of BSs takes place. In between, each RRH is connected to the BBU pool using a dedicated point-to-point FrontHaul (FH) link to transport time-domain data, standardized as Common Public Radio Interface (CPRI) transport protocol [3]. Hence, C-RAN has several merits, including a full Coordinated Multi-Point (CoMP) processing capability to boost spectral efficiency [4] and resource multiplexing gains at the BBU pool for scalable deployment [5].

Despite the above advantages, the excessive FH capacity requirement [6] has led to an overall revisit of the C-RAN prototype, and thus the notion of *functional splits* between the RRH and BBU are proposed [7]. In this sense, RRHs become active components that host a subset of network functionalities called (Remote) Radio Units (RUs/RRUs). The BBU handles the remaining network function processing, and then a disaggregated RAN architecture is formed[1]. In this regard, data are transported over the FH link according to the applied functional splits, such as the options defined by third Generation Partnership Project (3GPP) and by CPRI initiative as enhanced CPRI (eCPRI). This also creates an opportunity to transport Radio over Ethernet (RoE) as a

[1] BBU can be further decomposed into Distributed Unit (DU) and Centralized Unit (CU), and a three-tier architecture is formed: CU, DU, and RU.

cost-effective alternative [8]. In this sense, the point-to-point FH link can evolve into a multi-segment FH mesh network requiring extra routing and switching functionalities [9], and thus the extra multiplexing gains over the FH network are presented.

To exploit the above multiplexing benefits of BBU pooling and FH networking, we can partition all the RRUs into several RRU clusters [10]. In this regard, every RRU in the same cluster must transport its data to the same BBU in the BBU pool, which is called the *anchor BBU* of this RRU cluster. One important remark is that RRUs within the same cluster can be jointly processed for coordination, whereas RRUs belonging to different clusters only cooperate opportunistically. In this regard, end-users that are associated with different RRU clusters may have different service performances, for example, data throughput. Therefore, one challenge is to properly associate these users with the formed RRU clusters to strike a balance between the service performance and multiplexing benefits. In addition, different functional splits between the RRU and BBU also impact performance by applying different coordination schemes.

In summary, five design factors are considered together in this work to unleash the full potential of a disaggregated RAN deployment: (1) RRU clustering, (2) User association, (3) RAN functional split, (4) FH network routing, and (5) BBU placement. Note that these factors are tightly coupled and impact each other. For example, to retain better performance for associated end-users, we can form a large RRU cluster to coordinate processing from multiple RRUs at the cost of a large FH link capacity and a powerful anchor BBU processing capability. To counter this cost, a less-centralized function split can be applied between the RRU cluster and its anchor BBU. To the best of our knowledge, this is the first time a problem covering these five factors is formulated, and the corresponding solution is provided to be applicable to any disaggregated RAN deployment.

The remainder of this paper is organized as follows. A brief state-of-the-art review is presented in Section II. We then introduce our system model of a two-tier disaggregated RAN in Section III and formulate the problem in Section IV to maximize network spectral efficiency. This problem covers all five design factors that should be considered in a general disaggregated RAN deployment. Subsequently, in Section V, we propose a two-level turbo-based solution to address the reformulated problem. To show its effectiveness, the simulation results are provided in Section VI for small- and medium-scale network typologies. Finally, the extensions and applicability of the proposed solution is addressed in Section VII, and concluding remarks are presented in Section VIII.

II. BACKGROUND KNOWLEDGE AND RELATED WORK

In this section, we first outline the background knowledge of RAN disaggregation in terms of functional split and network topology, and then review several related works.

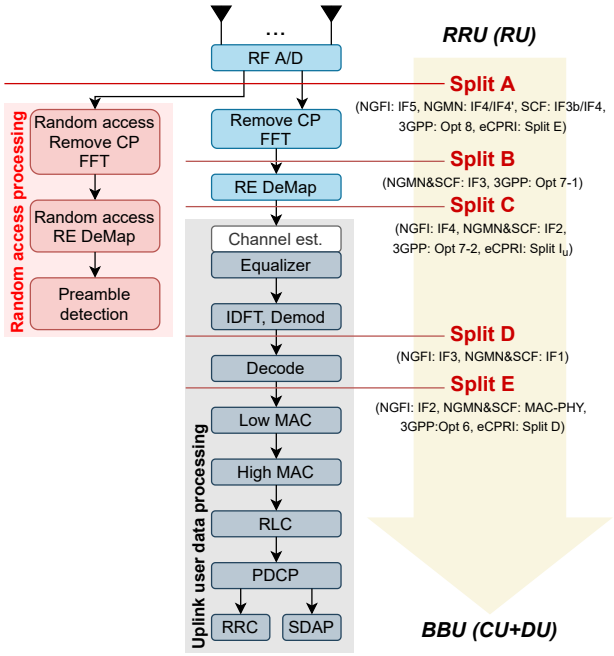


FIGURE 1: Considered RAN functional splits for uplink direction.

A. RAN FUNCTIONAL SPLIT

The RAN functional split between the RRU and the BBU affects (1) CoMP processing capability, (2) FH traffic data rate, and (3) network function execution time [11]. Because our focus is on data transportation over the FH network, only physical layer functional splits are considered, namely splits A, B, C, D, and E, as shown in Figure 1. Note that higher-layer functional splits (e.g., the one between the RLC and the PDCP) have a much more relaxed delay constraint than our considered lower-layer functional splits [12]. In addition, because uplink processing at the RRU can only be initiated after acquiring the air-interface signal, our focus in this work is on the uplink direction. In contrast, most downlink processing can be prepared beforehand [6]. It can also be observed in Figure 1 that our considered five functional splits can be mapped directly to those defined by Next Generation Fronthaul Interface (NGFI) [13], Small Cell Forum (SCF) [12], Next Generation Mobile Networks (NGMN) alliance [14], 3GPP [15], and eCPRI [16]. And we refer the interested readers to [17] for detailed elaborations.

B. NETWORK TOPOLOGY

The initial C-RAN topology features a dedicated point-to-point FH link between the RRU and the BBU. However, FH networks have evolved to support more complex topologies (e.g., tree, mesh) [18]. In this work, we focus on a multi-segment FH network that transports data in a two-tier disaggregated RAN. An example is depicted in Figure 2 with three RRU clusters and two BBU pools. First, a portion of the RAN processing, depending on the functional split, is performed at one type of disaggregated RAN node, that is, the RRU and BBU for the uplink and downlink directions,

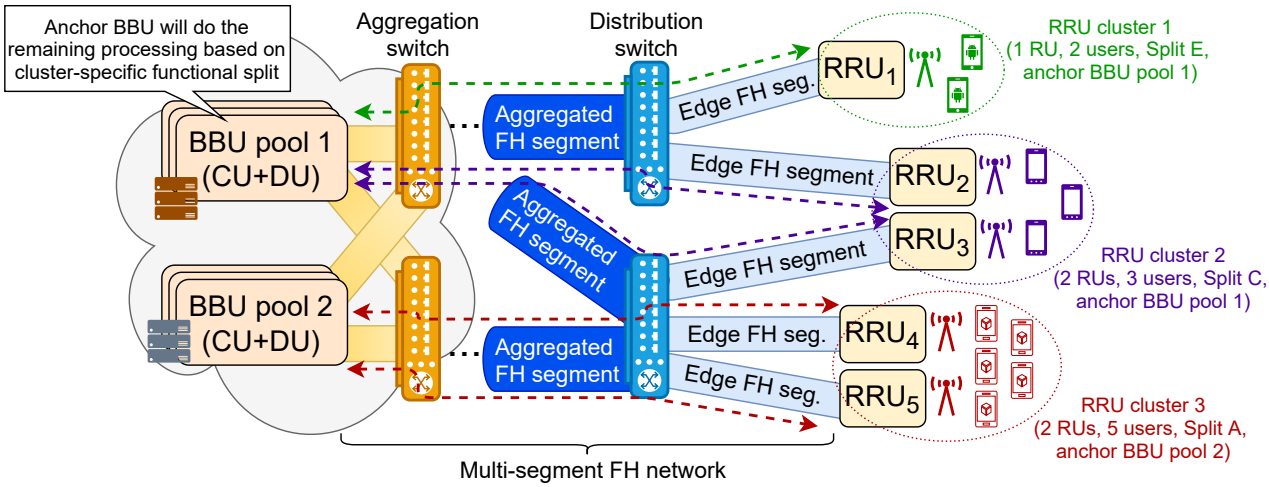


FIGURE 2: Network topology example: A two-tier disaggregated RAN over multi-segment FH network.

respectively. Then, the intermediate data are transported over the multi-segment FH network toward another type of disaggregated RAN node, that is, the BBU and RRU for the uplink and downlink directions, respectively. Finally, the remaining RAN processing is performed depending on the functional split. As previously mentioned, each BBU can be further decomposed into DU and CU deployed at different cloud locations.

Also in Figure 2, we can see that RRU clusters are formed to facilitate a joint CoMP processing, and thus RRUs within the same RRU cluster will apply the same functional split and be anchored to the same BBU. Note that this RRU clustering operation shall be executed dynamically [19] to deal with time-varying air-interface interference as well as transport FH traffic flows over the dynamic bandwidth allocated by Software-Defined Networking (SDN) controller. In addition, a particular group of end-users will be associated with the corresponding RRU cluster to take advantage of the CoMP processing within the cluster. It is worth noting that the user association problem is mutually dependent on the RRU clustering problem; therefore, they should be considered simultaneously.

C. RELATED WORK

In addition to the five design factors for a disaggregated RAN mentioned in Section I, we summarize some prior works addressing one or some of these factors as follows. The work in [20] formulates the RRH clustering problem as a bin packing problem and provides the heuristic solution to deal with the problem. Regarding the joint problem of user association and RRH clustering, the authors of [21] provided a sequential and heuristic solution based on the greedy algorithm, whereas the authors of [22] decomposes the original problem and provides an iterative solution. Moreover, a joint user association and RRH-BBU mapping problem was examined in [23], and a problem decomposition approach is applied. Furthermore, the authors of [24] jointly dealt with the factors of functional split and FH network routing to minimize RAN expenditure. In conclusion, the above studies investigated some of the aforementioned five design factors, and therefore can only be applied to specific disaggregated RAN deployments, for example, fixed functional split, static RRU clusters, dedicated FH transport networks, or predetermined BBU placement. By contrast, our work aims to place all five design factors in the same table and provide a unified approach to handle their interplay over general deployment.

Moreover, the authors of [25] established a framework to converge optical-wireless networks to minimize network deployment costs by considering a variety of factors, including server selection, transport network routing, and cross-domain resource allocation. Another study [26] investigated the joint problem of functional split, BBU allocation, and server scheduling, while minimizing the average end-to-end delay. In addition, several studies explored both RAN and Multi-access Edge Computing (MEC) domains simultaneously, among which LayBack [27] facilitated RAN communication and MEC computation resources into distinct layers, FluidRAN [28] jointly studied a virtualized RAN (vRAN)/MEC solution to minimize operational costs, and Matryoshka [29] tackled a multi-factor scheduling problem for computing resources, MEC services, and RAN workloads. Our work can be viewed as a complementary effort to the above works because it focuses on the design factors to be applied to a general disaggregated RAN deployment, and thus can provide control information from RAN deployment to other domains (e.g., SDN and MEC) or network services (e.g., bandwidth-guaranteed network slice).

TABLE 1: Parameter notation.

Parameter	Description
$\mathcal{R}$	The set of RRUs
$\mathcal{B}$	The set of BBUs
$\mathcal{U}$	The set of users
$\mathcal{C}$	The set of RRU clusters
$\mathcal{V}$	The set of RRUs, BBUs, and forwarding nodes in multi-segment FH network
$\mathcal{F}$	The set of functional splits
$\mathcal{E}$	The set of FH links between RRU, BBU, and forwarding node in multi-segment FH network
$\mathbf{Q}$	Detectable RRU indicator matrix
$\overline{\mathbf{Q}}$	Detectable RRU indicator matrix after RRU clustering
$\mathbf{C}$	RRU clustering variable matrix
$\overline{\mathbf{C}}$	Inter-RRU clustering relation indicator matrix
$\mathbf{N}$	RRU normalization diagonal matrix
$\mathbf{X}$	User association variable matrix
$\overline{\mathbf{X}}$	Inter-user interference indicator matrix
$\mathbf{F}$	Functional split variable matrix
$\mathbf{A}$	BBU placement variable matrix
$\mathbf{E}$	FH routing variable matrix
$\overline{\mathbf{E}}$	Routable FH link toward BBU indicator matrix
$\mathbf{h}_{i,j}$	Channel vector from the j-th user to the i-th RRU
$\mathcal{N}_j$	The set of all serving RRUs for the j-th user
$\mathbf{T}$	Signal to Interference plus Noise Ratio (SINR) matrix
Π	The set of composite variables in terms of functional split, BBU placement, and FH routing

III. SYSTEM MODEL

In this section, we elaborate on our system model considering all five design factors. The detailed parameter notation table[2] can be found in Table 1.

A. NETWORK TOPOLOGY

In our considered network topology, there are $|\mathcal{R}|$ RRUs $\mathcal{R} \coloneqq \{r_1, \cdots, r_{|\mathcal{R}|}\}$ (each with M antennas) serving all $|\mathcal{U}|$ users $\mathcal{U} \coloneqq \{u_1, \cdots, u_{|\mathcal{U}|}\}$ (each with a single antenna). The $M \times 1$ fading channel vector from the j-th user to the i-th RRU in the uplink direction at time t is denoted as $\mathbf{h}_{i,j}(t) \sim CN\left(\mathbb{0}_{M\times 1}, \sigma_{i,j}^2 \cdot \mathbf{I}_M\right)$, where $CN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate complex Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $\sigma_{i,j}^2$ is the channel gain variance from the j-th user to the i-th RRU. The time index t is dropped in the remainder of this work for simplicity. Moreover, all $|R|$ RRUs are connected to $|\mathcal{B}|$ centralized BBUs $\mathcal{B} \coloneqq \{b_1, \cdots, b_{|\mathcal{B}|}\}$ through a number of forwarding nodes $\mathcal{D} \coloneqq \{d_1, d_2, \cdots\}$ in a multi-segment FH network. Finally, all nodes in our considered network topology are denoted in set $\mathcal{V} \coloneqq \mathcal{R} \cup \mathcal{D} \cup \mathcal{B}$, and all directional FH links are denoted in set $\mathcal{E}$.

[2] Additionally, bold uppercase letters denote matrices; bold lowercase letters denote column vectors; $(\cdot)^T$ and $(\cdot)^H$ are transposition and Hermitian transposition operators, respectively; $\|\cdot\|_p$ is the p-norm of the column vector; $\mathbb{1}_{M\times N}$ and $\mathbb{0}_{M\times N}$ are the $M \times N$ all-ones and all-zeros matrices, respectively; $\mathbf{I}_N$ is the $N \times N$ identity matrix.

B. USER ASSOCIATION AND RRU CLUSTERING

For the j-th user, we can write its adjacent RRU as r_j and its detectable RRU set as R_j in Eq. (1a) and Eq. (1b), respectively. Note that γ_{th} in Eq. (1b) is the minimum signal power threshold required for a user to detect an RRU. Based on the above definition, we can form one $|\mathcal{R}| \times |\mathcal{U}|$ matrix $\mathbf{Q}$, in which its (i,j)-th element $q_{i,j}$ is 1 if $r_i \in R_j$ and is 0 otherwise.

$$r_j = \arg\max_{r_i \in \mathcal{R}} \sigma_{i,j}^2 \tag{1a}$$

$$R_j = \{r_j\} \cup \left\{r_i \in \mathcal{R} : \sigma_{i,j}^2 \geqslant \gamma_{th}\right\} \tag{1b}$$

Moreover, as mentioned in Section II-B, several RRU clusters are formed and denoted as set $\mathcal{C} \coloneqq \{c_1, \cdots, c_{|\mathcal{C}|}\}$, in which there are $|\mathcal{C}|$ RRU clusters and c_l contains all the RRUs within the l-th RRU cluster. Nevertheless, owing to the limited hosting capability of BBU, at most $C_{\max}$ RRU can be coordinated at a time, i.e., $|c_l| \leqslant C_{\max}, \forall c_l \in \mathcal{C}$. Furthermore, we can represent such RRU clustering in matrix form to simplify notation; therefore, one $|\mathcal{R}| \times |\mathcal{R}|$ matrix is denoted as $\mathbf{C}$, in which its (i,j)-th element $c_{i,j}$ is 1 if the i-th RRU belongs to the j-th RRU cluster and is 0 otherwise. Thanks to the above matrix representation, one $|\mathcal{R}| \times |\mathcal{R}|$ inter-RRU clustering relation matrix can be directly written as $\overline{\mathbf{C}} = \mathbf{C} \cdot \mathbf{C}^T$, where its (i,j)-th element $\overline{c}_{i,j}$ is 1 if both the i-th and the j-th RRUs belong to the same RRU cluster and is 0 otherwise. Finally, one $|\mathcal{R}| \times |\mathcal{R}|$ RRU normalization matrix is expressed as $\mathbf{N}$ in Eq. (2), where the diag $(\cdot)$ operator can create a diagonal matrix whose diagonal entries are given by the entries of the vector.

$$\begin{aligned}\mathbf{N} &= \begin{bmatrix} n_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & n_{|\mathcal{R}|} \end{bmatrix} \\ &= \text{diag}\left(\left[\frac{1}{\sum_{j=1}^{|\mathcal{R}|} \overline{c}_{1,j}}, \cdots, \frac{1}{\sum_{j=1}^{|\mathcal{R}|} \overline{c}_{|\mathcal{R}|,j}}\right]\right)\end{aligned} \tag{2}$$

Based on the above notations for user association and RRU clustering, we take an additional step in defining extra notations. First, since all RRUs are now grouped into RRU clusters, we can extend the above detectable RRU matrix $\mathbf{Q}$ directly as $\overline{\mathbf{Q}} \coloneqq \overline{\mathbf{C}} \cdot \mathbf{Q} \succ \mathbb{0}_{|R|\times|U|}$, in which $\succ$ is the element-wise "greater than" relational operator and its (i,j)-th element $\overline{q}_{i,j}$ is 1 if the j-th user can detect, after considering RRU clustering, the i-th RRU and is 0 otherwise. Second, one $|\mathcal{R}| \times |\mathcal{U}|$ variable matrix $\mathbf{X}$ is used to represent the user association decision, in which its (i,j)-th element $x_{i,j}$ is 1 if the j-th user is associated with the i-th RRU and is 0 otherwise. Finally, the inter-user interference matrix is denoted as $\overline{\mathbf{X}} \coloneqq \mathbf{X}^T \cdot \left(\mathbb{1}_{|R|\times|U|} - \mathbf{X}\right) \succ \mathbb{0}_{|U|\times|U|}$, where its (i,j)-th element $\overline{x}_{i,j}$ is 1 if interference exists between the i-th user and the j-th user and is 0 otherwise.

C. FUNCTIONAL SPLIT AND BBU PLACEMENT

As mentioned in Section II-A, the RAN processing is decomposed into several network functions to be placed based

TABLE 2: Per-link parameters of multi-segment FH network.

Characteristic	Parameter	Description
Delay	$\mathbf{t}$	FH link delay in seconds
Capacity	$\mathbf{l}$	FH link capacity in bits per second

on the applied *functional split*. We first denote the set of considered functional splits as $\mathcal{F} \coloneqq \{f_1, \cdots, f_{|F|}\}$, i.e., it contains the five splits shown in Figure 1. Moreover, one $|\mathcal{R}| \times |\mathcal{F}|$ variable matrix $\mathbf{F}$ is defined to represent the functional split decision, in which its (i,j)-th element $f_{i,j}$ is 1 if the i-th RRU applies the j-th functional split and is 0 otherwise. Furthermore, one $|\mathcal{R}| \times |\mathcal{B}|$ BBU placement variable matrix $\mathbf{A}$ is defined, in which its (i,j)-th element $a_{i,j}$ is 1 if the i-th RRU places its remaining processing at the j-th BBU and is 0 otherwise. In addition, there are two specific remarks. First, different functional splits result in different CoMP schemes, which will be elaborated in detail in Section IV. Second, to enable a CoMP scheme for all RRUs within the same RRU cluster, they shall apply an identical functional split and place their remaining processing at the same anchor BBU.

D. MULTI-SEGMENT FRONTHAUL ROUTING

To model the delay and capacity of the FH link in the multi-segment FH network, i.e., $\mathcal{E}$, two parameters $\mathbf{t}$ and $\mathbf{l}$ are denoted in Table 2. They are both $|\mathcal{E}| \times 1$ column vectors, and their i-th entry, i.e., t_i and l_i, correspond to the i-th FH link within $\mathcal{E}$. Moreover, we define one $|\mathcal{R}| \times |\mathcal{E}|$ variable matrix $\mathbf{E}$, in which its (i,j)-th entry $e_{i,j}$ is 1 if the j-th FH link in $\mathcal{E}$ is decided to route the FH traffic from the i-th RRU to its anchor BBU and is 0 otherwise. Finally, one $|\mathcal{E}| \times |\mathcal{B}|$ routable FH link matrix $\overline{\mathbf{E}}$ is defined, in which its (i,j)-th entry $\overline{e}_{i,j}$ is 1 if the i-th FH link in $\mathcal{E}$ can be used (by any RRU) to route any FH traffic to the j-th BBU and is 0 otherwise.

IV. PROBLEM FORMULATION AND ANALYSIS

In this section, we first formulate the overall problem to maximize network spectral efficiency, and then explain each constraint. Subsequently, the expected SINR of different CoMP schemes is derived in their respective closed forms for five different functional splits. Finally, a complexity analysis of the formulated problem is performed.

A. OBJECTIVE FUNCTION

The overall problem is formulated in Eq. (3), and its objective function in Eq. (3a) aims to maximize the network spectral efficiency summed from all users while still satisfying a number of constraints from Eq. (3b) to Eq. (3p). In specific, R_j denotes the network spectral efficiency in bits per second per Hertz (bps/Hz) experienced by the j-th user. Note that this is derived based on the Shannon capacity formula, in which $E[\tau_{j,m}]$ is the expected SINR of the j-th user when applying the m-th functional split, and x_{i_j}, n_i, and $f_{i,m}$ have already been introduced in Sections III-B and III-C. To align with several matrix forms denoted in Section III, this objective function can be further describe in a more compact format. In detail, one $|U| \times |F|$ matrix $\mathbf{T}$ is defined in Eq. (4) to include all SINR $\tau_{j,m}$ of every user and every functional split. Afterwards, we apply the entry-wise Hadamard product operator (i.e., $\circ$ in Eq. (3a)) to realize the entry-for-entry product of two equally-sized matrices.

$$\mathbf{T} = \begin{bmatrix} \tau_{1,A} & \tau_{1,B} & \tau_{1,C} & \tau_{1,D} & \tau_{1,E} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \tau_{|U|,A} & \tau_{|U|,B} & \tau_{|U|,C} & \tau_{|U|,D} & \tau_{|U|,E} \end{bmatrix} \tag{4}$$

B. PROBLEM CONSTRAINTS

Within the formulated problem, we can classify all constraints into four categories, and then interpret each category respectively in the following paragraphs.

1) User association and RRU clustering

There are five constraints belonging to this category, from Eq. (3b) to Eq. (3f). Among these, Eq. (3b) is the most intuitive and can guarantee that each RRU belong to one RRU cluster. According to Eq. (3c), the number of RRUs within an RRU cluster shall not exceed $C_{\max}$, as mentioned in Section III-B. Moreover, Eq. (3d) aims to prevent each user from being associated with multiple RRU clusters. Besides, Eq. (3e) ensures that each user is only associated with its detectable RRUs after considering RRU clustering, based on $\overline{q}_{i,j}$ within $\overline{\mathbf{Q}}$ defined in Section III-B. Finally, Eq. (3f) provides several insights, as explained in Table 3, by exploiting the property of symmetric matrix $\overline{\mathbf{C}}$, i.e., $\overline{c}_{i,k} = \overline{c}_{k,i}, \forall i,k$.

2) Functional split

There are two constraints fall into this category, Eq. (3g) and Eq. (3h). In Eq. (3g), each RRU is restricted to using only one functional split within the set $\mathcal{F}$. Moreover, Eq. (3h) guarantees that the RRUs in the same RRU cluster apply an identical functional split, as mentioned in Section III-C. In specific, Table 4 explains all feasible combinations.

3) BBU placement and FH network routing

There are five constraints in this category, from Eq. (3i) to Eq. (3m). First, Eq. (3i) ensures that each RRU is anchored to a single BBU. Then, Eq. (3j) guarantees that RRUs in the same RRU cluster are anchored to the same BBU, and all its feasible combinations are presented in Table 5. To proceed one step forward, two functions are defined for each node $v \in \mathcal{V}$ in the network topology: $\delta^+(v)$ and $\delta^-(v)$ represent the outgoing and incoming FH links of node v, respectively. Based on these two functions, the standard flow conservation constraint is given in Eq. (3k) from each source node (i.e., RRU) to its sink node (i.e., anchor BBU). In addition, Eq. (3l) ensures that the outgoing degree of each node cannot be greater than 1 to avoid an unwanted routing loop. Finally, Eq. (3m) avoids using FH links that cannot be routed from each source node to its sink node, by leveraging previously defined $\overline{e}_{\epsilon,j}$ within $\overline{\mathbf{E}}$ in Section III-D.

TABLE 3: Possible combinations of Eq. (3f).

$x_{i,j}$	$\overline{c}_{i,k}$	$x_{k,j}$	**Description**
1	1	1	u_j is associated with r_i & r_i and r_k belong to the same RRU cluster $\rightarrow$ u_j is associated with r_k
0	1	0	u_j is not associated with r_i & r_i and r_k belong to the same RRU cluster $\rightarrow$ u_j is not associated with r_k
1	0	0	u_j is associated with r_i & r_i and r_k belong to different RRU clusters $\rightarrow$ u_j is not associated with r_k
0	0	0 or 1	u_j is not associate with u_j & r_i and r_k belong to different RRU clusters $\rightarrow$ u_j may or may not be associated with r_k

TABLE 4: Possible combinations of Eq. (3h).

$f_{i,m}$	$\overline{c}_{i,k}$	$f_{k,m}$	**Description**
1	1	1	r_i uses the m-th functional split & r_i and r_k belong to the same RRU cluster $\rightarrow$ r_k uses the m-th functional split
0	1	0	r_i does not use the m-th functional split & r_i and r_k belong to the same RRU cluster $\rightarrow$ x_i does not use the m-th functional split
1	0	0 or 1	r_i and r_k belong to different RRU clusters $\rightarrow$ r_i and r_k may or may not use the same functional split
0	0		

TABLE 5: Possible combinations of Eq. (3j).

$a_{i,n}$	$\overline{c}_{i,k}$	$a_{k,n}$	**Description**
1	1	1	r_i is anchored to b_n & r_i and r_k belong to the same RRU cluster $\rightarrow$ r_k is anchored to b_n
0	1	0	r_i is not anchored to b_n & r_i and r_k belong to the same RRU cluster $\rightarrow$ r_k is not anchored to b_n
1	0	0 or 1	r_i and r_k belong to different RRU clusters $\rightarrow$ r_i and r_k may or may not be anchored to the same BBU
0	0		

$$\underset{\mathbf{C},\mathbf{X},\mathbf{F},\mathbf{A},\mathbf{E}}{\text{maximize}} \sum_{u_j\in\mathcal{U}} R_j = \sum_{u_j\in\mathcal{U}}\sum_{r_i\in\mathcal{R}}\sum_{f_m\in\mathcal{F}} x_{i,j}\cdot n_i\cdot f_{i,m}\cdot\log\left(1+E\left[\tau_{j,m}\right]\right) = \mathbb{1}_{|\mathcal{U}|\times 1}^{T}\cdot\left(\mathbf{X}^T\cdot\mathbf{N}\cdot\mathbf{F}\circ\log\left(\mathbb{1}_{|\mathcal{U}|\times|\mathcal{F}|}+E\left[\mathbf{T}\right]\right)\right)\cdot\mathbb{1}_{|\mathcal{F}|\times 1} \tag{3a}$$

$$\text{subject to}\quad \mathbf{C}\cdot\mathbb{1}_{|\mathcal{R}|\times 1}=\mathbb{1}_{|\mathcal{R}|\times 1} \tag{3b}$$

$$\left\|\mathbf{C}^T\cdot\mathbb{1}_{|\mathcal{R}|\times 1}\right\|_{\infty}\leqslant C_{\max} \tag{3c}$$

$$\mathbf{X}^T\cdot\mathbf{N}\cdot\mathbb{1}_{|\mathcal{R}|\times 1}=\mathbb{1}_{|\mathcal{U}|\times 1} \tag{3d}$$

$$x_{i,j}\leqslant\overline{q}_{i,j},\quad \forall r_i\in\mathcal{R},\ u_j\in\mathcal{U} \tag{3e}$$

$$x_{i,j}\cdot\overline{c}_{i,k}=x_{k,j}\cdot\overline{c}_{k,i}=x_{i,j}\cdot x_{k,j},\quad \forall r_i\neq r_k\in\mathcal{R},\ u_j\in\mathcal{U} \tag{3f}$$

$$\mathbf{F}\cdot\mathbb{1}_{|\mathcal{F}|\times 1}=\mathbb{1}_{|\mathcal{R}|\times 1} \tag{3g}$$

$$f_{i,m}\cdot\overline{c}_{i,k}=f_{k,m}\cdot\overline{c}_{k,i},\quad \forall r_i\neq r_k\in\mathcal{R},\ f_m\in\mathcal{F} \tag{3h}$$

$$\mathbf{A}\cdot\mathbb{1}_{|\mathcal{B}|\times 1}=\mathbb{1}_{|\mathcal{R}|\times 1} \tag{3i}$$

$$a_{i,n}\cdot\overline{c}_{i,k}=a_{k,n}\cdot\overline{c}_{k,i},\quad \forall r_i\neq r_k\in\mathcal{R},\ b_n\in\mathcal{B} \tag{3j}$$

$$\sum_{\epsilon\in\delta^+(v)} e_{i,\epsilon}-\sum_{\epsilon\in\delta^-(v)} e_{i,\epsilon}=\begin{cases}-1, & \text{if } v=r_i\\ a_{i,n}, & \text{if } v=b_n\in\mathcal{B}\\ 0, & \text{else}\end{cases},\forall r_i\in\mathcal{R},\ v\in\mathcal{V} \tag{3k}$$

$$\sum_{\epsilon\in\delta^+(v)} e_{i,\epsilon}\leqslant 1,\quad \forall r_i\in\mathcal{R},\ v\in\mathcal{V} \tag{3l}$$

$$a_{i,n}\cdot e_{i,\epsilon}\leqslant\overline{e}_{\epsilon,n},\quad \forall r_i\in\mathcal{R},\ b_n\in\mathcal{B},\ \epsilon\in\mathcal{E} \tag{3m}$$

$$T_R\left(f_{i,1},\cdots,f_{i,|\mathcal{F}|}\right)+\sum_{e_{i,\epsilon}==1,\forall\epsilon\in\mathcal{E}} t_\epsilon+T_B\left(f_{i,1},\cdots,f_{i,|\mathcal{F}|}\right)\leqslant T_{rx}^{\max},\forall r_i\in\mathcal{R} \tag{3n}$$

$$\sum_{r_i\in\mathcal{R}} e_{i,\epsilon}\cdot W_R\left(f_{i,1},\cdots,f_{i,|\mathcal{F}|},x_{i,1},\cdots,x_{i,|\mathcal{U}|}\right)\leqslant l_\epsilon,\forall\epsilon\in\mathcal{E} \tag{3o}$$

$$x_{i,j},c_{i,k},f_{i,m},a_{i,n},e_{i,\epsilon}\in\{0,1\},\forall r_i,r_k\in\mathcal{R},x_j\in\mathcal{U},f_m\in\mathcal{F},b_n\in\mathcal{B},\epsilon\in\mathcal{E} \tag{3p}$$

4) FH traffic transportation

Two constraints related to this category, i.e., Eq. (3n) and Eq. (3o), are added to ensure that the FH traffic from all RRUs can be accommodated in the multi-segment FH network. First, Eq. (3n) guarantees that the FH traffic from each RRU to its anchor BBU does not violate the maximum delay allowed for uplink reception, i.e., $T_{rx}^{\max}$. In specific, there are three components on the left-hand side of Eq. (3n): (1) RRU processing time $T_R(\cdot)$, (2) summation of per-link delay t_ϵ over the FH routing path, and (3) BBU processing time $T_B(\cdot)$. It is worth noting that both the RRU processing time and BBU processing time depend on the applied functional split, i.e., $f_{i,m}$ for the i-th RRU, and can be measured using a known framework, e.g., OpenAirInterface.

The next constraint in Eq. (3o) ensures that the per-link capacity l_ϵ will not be exceeded by all FH traffic. In this realization, the left-hand side of Eq. (3o) takes the summation of FH datarate, i.e., $W_R(\cdot)$, from every RRU that utilizes this link (i.e., $e_{i,\epsilon}$ is 1) to route its FH traffic. We notice that the FH datarate depends not only on the applied functional split (e.g., $f_{i,m}$ for the i-th RRU) but also on the associated users (e.g., $x_{i,j}$ for the i-th RRU). To model it numerically, we apply the same approach as in our previous work [6].

C. EXPECTED SINR FORMULATION

In addition to the above constraints, we formulate the expected SINR in the objective function, i.e., $E[\tau_{j,m}]$, into their respective closed forms. To facilitate our derivations, we first denote the power of transmitted symbols from each user and the power of Additive White Gaussian Noise (AWGN) as $P_s = 1$ and N_0, respectively. Also, we concatenate all fading channel vectors (cf. Section III-A) from the j-th user to every RRU, i.e., $\mathbf{h}_{i,j}\ \forall r_i \in \mathcal{R}$, and build one aggregated channel vector $\mathbf{h}_j$ for the j-th user in Eq. (5).

$$\mathbf{h}_j = \begin{bmatrix} \mathbf{h}_{1,j} \\ \vdots \\ \mathbf{h}_{|R|,j} \end{bmatrix} \tag{5}$$

In the following, we first introduce the applicable CoMP schemes and then formulate the corresponding SINR form of each functional split, i.e., from $\tau_{j,A}$ to $\tau_{j,E}$. Finally, the expected SINR will be derived correspondingly. In particular, three CoMP schemes are considered: (a) **Joint reception** over time, frequency, and user domains for split A, split B, and split C, respectively, (b) **Soft symbol combination** for split D, and (c) **Transport block selection** for split E.

1) Joint reception

Between the time and frequency domains, there is no performance difference when performing joint reception because they can be transformed interchangeably using (Inverse) Discrete Fourier Transformation (DFT/IDFT) operations. In this regard, the corresponding SINR forms of split A and split B (i.e., $\tau_{j,A}$ and $\tau_{j,B}$) can be written in Eq. (6a) by applying Minimum Mean Square Error (MMSE) receiver vector $\mathbf{w}_j$ and following the derivation steps introduced in Appendix A-A (from Eq. (13) to Eq. (12c)).

In comparison, the joint reception over the user domain can only process the received uplink signal independently among user-specific resource blocks; therefore, the Inter-Carrier Interference (ICI) [30] produced by different users can deteriorate the performance and decrease SINR[3]. To quantitatively model this interference, a simple approach from [31] is applied, and a portion of the transmitted power, i.e., $0 \leqslant r_{ici} \leqslant 1$, is treated as interference. Therefore, we can follow almost the same derivation approach introduced in Appendix A-A (i.e., from Eq. (14a) to Eq. (14c)) and apply the same MMSE principle to obtain the corresponding SINR form of split C $\tau_{j,C}$ in Eq. (6b). We notice that this SINR form of split C will be the same as $\tau_{j,A}$ and $\tau_{j,B}$ in Eq. (6a) when r_{ici} is zero, i.e., no ICI between users.

2) Soft symbol combination

This CoMP scheme aims to combine the processed symbols from different RRUs at the anchor BBU. In specific, each RRU perform the bsaeband processing until the demodulation network function (cf. Figure 1) and then transport "soft" symbols[4] over the FH network to be combined by the anchor BBU. Therefore, such a scheme is suitable for split D.

To derive the corresponding SINR form, we first write the SINR of the soft symbols from the j-th user to the i-th RRU as $\tau_{i,j}^{ss}$ in Eq. (6c) (Refer to Appendix A-B for a detailed derivation). The SINR of soft symbols can be viewed as the signal quality after RRU processing. Then, to achieve the maximum SINR after combination, we apply the Maximal Ratio Combining (MRC) approach [32] at the anchor BBU to combine soft symbols from all RRUs in the same RRU cluster. To this end, the SINR form of split D $\tau_{j,D}$ can be written as the summation of the SINR of all soft symbols $\tau_{i,j}^{ss}$ in Eq. (6d). Note that such summation is done over all RRUs in the same RRU cluster, i.e., $x_{i,j}$ equals 1 for the j-th user, and thus we define a new set $\Gamma_j = \{r_i : x_{i,j} == 1\}$ including all serving RRUs for the j-th user.

3) Transport block selection

Unlike other CoMP schemes, this method can select only the successfully received transport block that passed the Cyclic Redundancy Check (CRC) from different RRUs. In this regard, each RRU is responsible for all physical layer processing until the end of the channel decoder (cf. Figure 1); thus, the anchor BBU only performs layer 2 and above processing. We can observe that this scheme is a good match to split E. Finally, the SINR form of split E $\tau_{j,E}$ is viewed as selecting the maximum SINR among all the soft symbols in Eq. (6e), in which $\tau_{i,j}^{ss}$ and Γ_j are introduced in Eq. (6c) and Eq. (6d), respectively.

[3] Such an ICI is due to unequal carrier frequency offsets between users, and its main root causes are oscillator mismatches and user mobility.

[4] In comparison, hard symbols refer to quantized constellation points according to the modulation scheme allocated to each user.

4) Expected SINR

Based on the above SINR forms in Eq. (6), we further derive their respective expected values, i.e., from $E[\tau_{j,A}]$ to $E[\tau_{j,E}]$ in Eq. (7), to be used in the objective function.

The expected SINRs of splits A and B are formulated in Eq. (7a) by following the derivation in Appendix A-C (from Eq. (17a) to Eq. (17c)). One can notice that such expected SINR is made up of three components: (1) the eigenvector-projected channel power ($\check{\sigma}^2_{j,k}$), (2) the noise variance (N_0), and (3) the joint Probability Density Function (PDF) of all eigenvalues from interfering users, i.e., $f\left(\lambda_{j,1}, \cdots, \lambda_{j,|\mathcal{N}_j|}\right)$. Note that the PDF of all eigenvalues can be derived using either random matrix theory [33] for some special forms or generated stochastically.

In addition, the expected SINR of split C is derived in Eq. (7b) by exploiting a similar approach as that of splits A and B. Due to the extra ICI, additional terms are added to model the interference, such as $z_{j,k}$, and we refer readers to Appendix A-C for more details. Besides, the ICI power ratio r_{ici} previously defined in Section IV-C1 also deteriorates the expected SINR by reducing the channel power by a factor of $(1 - r_{ici})$. We can see that the expected SINR of split C is the same as $E[\tau_{j,A}]$ and $E[\tau_{j,B}]$ in Eq. (7a) when there is no ICI between users (i.e., $r_{ici} = z_{j,k} = 0$).

Before deriving the expected SINR of splits D and E, we follow the approach in Appendix A-C and write the expected SINR of the soft symbols in Eq. (7c). Then, the expected SINRs of splits D and E are respectively derived in Eq. (7d) and Eq. (7e). After some inspections, we notice that the expected SINR of split D will is the same as $E[\tau_{j,A}]$ and $E[\tau_{j,B}]$ in Eq. (7a) if there is only one RRU in the RRU cluster or if there is no interfering user. Otherwise, the expected SINR of split D will be lower because the joint reception CoMP scheme will have more receiving antennas at the anchor BBU to reduce the impact of interference. Finally, the expected SINR of split E will be lower than that of split D, unless there is only one RRU in the RRU cluster.

$$\tau_{j,A} = \tau_{j,B} = \frac{\mathbf{w}_j^H \cdot \tilde{\mathbf{h}}_{j,j} \cdot \tilde{\mathbf{h}}_{j,j}^H \cdot \mathbf{w}_j}{\mathbf{w}_j^H \cdot \left(\sum_{u_k \in \mathcal{N}_j} \tilde{\mathbf{h}}_{j,k} \cdot \tilde{\mathbf{h}}_{j,k}^H + N_0 \cdot \mathbf{I}_{M_j} \right) \cdot \mathbf{w}_j} \tag{6a}$$

$$\tau_{j,C} = \frac{(1 - r_{ici}) \cdot \mathbf{w}_{j,C}^H \cdot \tilde{\mathbf{h}}_{j,j} \cdot \tilde{\mathbf{h}}_{j,j}^H \cdot \mathbf{w}_{j,C}}{\mathbf{w}_{j,C}^H \cdot \left(\sum_{u_k \in \mathcal{N}_j} \tilde{\mathbf{h}}_{j,k} \cdot \tilde{\mathbf{h}}_{j,k}^H + r_{ici} \cdot E\left[\tilde{\mathbf{h}}_{j,j} \cdot \tilde{\mathbf{h}}_{j,j}^H \right] + N_0 \cdot \mathbf{I}_{M_j} \right) \cdot \mathbf{w}_{j,C}} \tag{6b}$$

$$\tau_{i,j}^{ss} = \frac{\mathbf{w}_{i,j}^H \cdot \mathbf{h}_{i,j} \cdot \mathbf{h}_{i,j}^H \cdot \mathbf{w}_{i,j}}{\mathbf{w}_{i,j}^H \cdot \left(\sum_{u_k \in \mathcal{N}_j} \mathbf{h}_{i,k} \cdot \mathbf{h}_{i,k}^H + N_0 \cdot \mathbf{I}_M \right) \cdot \mathbf{w}_{i,j}} \tag{6c}$$

$$\tau_{j,D} = \sum_{r_i \in \mathcal{R}} x_{i,j} \cdot \tau_{i,j}^{ss} = \sum_{r_i \in \Gamma_j} \tau_{i,j}^{ss} \tag{6d}$$

$$\tau_{j,E} = \max_{r_i \in \Gamma_j} \tau_{i,j}^{ss} \tag{6e}$$

$$E[\tau_{j,A}] = E[\tau_{j,B}] = \sum_{k=1}^{|\mathcal{N}_j|} E\left[\frac{1}{\lambda_{j,k} + N_0} \right] \cdot \check{\sigma}^2_{j,k} + \sum_{k=|\mathcal{N}_j|+1}^{M_j} \frac{\check{\sigma}^2_{j,k}}{N_0} \tag{7a}$$

$$E[\tau_{j,C}] = \sum_{k=1}^{|\mathcal{N}_j|} E\left[\frac{1}{\lambda_{j,k} + z_{j,k} + N_0} \right] \cdot (1 - r_{ici}) \cdot \check{\sigma}^2_{j,k} + \sum_{k=|\mathcal{N}_j|+1}^{M_j} \frac{(1 - r_{ici}) \cdot \check{\sigma}^2_{j,k}}{z_{j,k} + N_0} \tag{7b}$$

$$E\left[\tau_{i,j}^{ss}\right] = \sigma^2_{i,j} \cdot \left(\sum_{k=1}^{|\mathcal{N}_j|} E\left[\frac{1}{\lambda^{ss}_{i,j,k} + N_0} \right] + \frac{M - |\mathcal{N}_j|}{N_0} \right) \tag{7c}$$

$$E[\tau_{j,D}] = \sum_{r_i \in \Gamma_j} \sigma^2_{i,j} \cdot \left(\sum_{k=1}^{|\mathcal{N}_j|} E\left[\frac{1}{\lambda^{ss}_{i,j,k} + N_0} \right] + \frac{M - |\mathcal{N}_j|}{N_0} \right) \tag{7d}$$

$$E[\tau_{j,E}] = \max_{r_i \in \Gamma_j} \sigma^2_{i,j} \cdot \left(\sum_{k=1}^{|\mathcal{N}_j|} E\left[\frac{1}{\lambda^{ss}_{i,j,k} + N_0} \right] + \frac{M - |\mathcal{N}_j|}{N_0} \right) \tag{7e}$$

D. PROBLEM ANALYSIS

To analyze the complexity of the overall problem in Eq. (3), we observe from the problem constraint in Eq. (3p) that the binary values (0 or 1) need to be assigned to all $|\mathcal{R}| \times (|\mathcal{R}| + |\mathcal{U}| + |\mathcal{F}| + |\mathcal{B}| + |\mathcal{E}|)$ variables belonging to five different variable matrices: RRU clustering ($c_{i,k}$ in $\mathbf{C}$), user association ($x_{i,j}$ in $\mathbf{X}$), functional split ($f_{i,m}$ in $\mathbf{F}$), BBU placement ($a_{i,n}$ in $\mathbf{A}$), and FH routing ($e_{i,\epsilon}$ in $\mathbf{E}$). Specifically, this problem can be proven to have *NP-hard* complexity.

Theorem IV.1. *The problem of Eq. (3) is NP-hard to solve.*

Proof. This problem can be polynomially reduced to a known NP-hard Multi-dimensional Multiple-choice Knapsack Problem (MMKP) [34]. We consider a specific instance of this problem with the following four characteristics. In the first place, such instance fixes RRU clustering ($\mathbf{C}$), user association ($\mathbf{X}$), and BBU anchoring ($\mathbf{A}$), while still satisfying the constraints from Eq. (3b) to Eq. (3f), Eq. (3i), and Eq. (3j). Second, its multi-segment FH network can guarantee that at least one feasible routing path exists between each pair of RRU and BBU that meet the constraints from Eq. (3k) to Eq. (3m). Third, all FH links have a sufficiently large capacity (e.g., $t_\epsilon \rightarrow \infty, \forall \epsilon \in \mathcal{E}$) and a negligibly small delay (e.g., $l_\epsilon \rightarrow 0^+, \forall \epsilon$); therefore, they do not affect on the constraints in Eq. (3n) and Eq. (3o). Finally, linear functions are applied to model $T_R(\cdot)$, $T_B(\cdot)$ in Eq. (3n), and $W_R(\cdot)$ in Eq. (3o). In the next step, we can rewrite all the remaining constraints, i.e., Eq. (3g), Eq. (3h), Eq. (3n), and Eq. (3o), as the linear functions of a new composite variable $\overline{f}_{j,m} = \sum_{r_i \in \mathcal{R}} x_{i,j} \cdot f_{i,m}, \forall u_j \in \mathcal{U}, f_m \in \mathcal{F}$. Additionally, the objective function can also be rewritten as a linear function of $\overline{f}_{j,m}$. To conclude, this specific problem instance can be mapped into one MMKP with $|\mathcal{U}|$ classes of items, and exactly one item is selected from $|\mathcal{F}|$ items for each individual class. □

V. PROBLEM REFORMULATION AND PROPOSED SOLUTION

In this section, to deal with the origianl problem in Eq. (3), we first reformulate its objective function and constraints into Eq. (8). Then, a two-level turbo-based solution is proposed to address the reformulated problem efficiently.

A. PROBLEM REFORMULATION

The original objective function in Eq. (3a) is related to both functional split ($\mathbf{F}$) and user association ($\mathbf{X}$). However, it is only linear in terms of $\mathbf{F}$ because the expected SINR forms in Eq. (7) are neither convex nor concave for $\mathbf{X}$. Therefore, one possible approach to such a situation is to iteratively update the expected SINR $E[\mathbf{T}]$ using the previous iteration $\mathbf{X}$ values. In this sense, the objective function can be viewed as a *bilinear function* for the continuously-relaxed $\mathbf{F}$ and $\mathbf{X}$, i.e., $f_{i,m}, x_{i,j} \in [0,1]$. In practice, a bilinear problem can be solved through which two sets of variables are tackled alternatively while fixing another set of variables, termed as alternate convex search [35]. Note that the continuously-relaxed $\mathbf{X}$ reshapes the original constraint in Eq. (3f) into the constraint $\mathcal{C}_3$ in Eq. (8d).

Moreover, by inspecting the constraints from Eq. (3g) to Eq. (3j), we notice that both functional split ($\mathbf{F}$) and BBU placement ($\mathbf{A}$) are highly related to one another. Therefore, they can be handled together using a composite variable $\phi^m_{i,n} = f_{i,m} \cdot a_{i,n}$ to indicate whether the i-th RRU uses the m-th functional split to transport its FH traffic toward the n-th BBU. Note that such a composite variable can also be relaxed continuously between 0 and 1, i.e., $\phi^m_{i,n} \in [0,1]$.

To go one step further by inspecting the constraints from Eq. (3k) to Eq. (3n), we can notice that the above composite variable $\phi^m_{i,n}$ significantly impacts the feasible FH routing paths. Take Eq. (3n) as an example, the feasible FH routing paths are obviously limited when both functional split and BBU placement are fixed. Under this circumstance, we can form a feasible FH routing path set $\mathcal{P}^m_{i,n}$ when the i-th RRU uses the m-th functional split to route its FH traffic to the n-th BBU (i.e., $\phi^m_{i,n}$ = 1). In specific, five FH routing path sets will be formed for each pair of RRU and BBU: $\mathcal{P}^A_{i,n} \subseteq \mathcal{P}^B_{i,n} \subseteq \mathcal{P}^C_{i,n} \subseteq \mathcal{P}^D_{i,n} \subseteq \mathcal{P}^E_{i,n}, \forall r_i \in \mathcal{R}, b_n \in \mathcal{B}$. Afterward, to represent the usage of the q-th FH routing path within set $\mathcal{P}^m_{i,n}$, we define another composite variable $\pi^m_{i,n,q}$, identifying whether this path is selected to deliver the FH traffic from the i-th RRU to the n-th BBU when using the m-th functional split. Finally, the relationship between the two defined composite variables, i.e., $\phi^m_{i,n}$ and $\pi^m_{i,n,q}$, can be found in Eq. (9).

$$\phi^m_{i,n} = f_{i,m} \cdot a_{i,n} = \sum_{q=1}^{\left|\mathcal{P}^m_{i,n}\right|} \pi^m_{i,n,q} \tag{9}$$

Based on our latest composite variable $\pi^m_{i,n,q}$, which jointly considers (a) functional split, (b) BBU placement, and (c) FH routing, we reformulate the constraints as follows:

1) $\mathcal{C}_4$ in Eq. (8e) replaces the constraints in Eq. (3g) and Eq. (3i) for each RRU,
2) $\mathcal{C}_5$ in Eq. (8f) replaces the constraints in Eq. (3h) and Eq. (3j) for each RRU, and
3) $\mathcal{C}_6$ of Eq. (8h) preserve the per-link capacity constraint in Eq. (3o) using an indicator function $I\left(\epsilon, P^m_{i,n,q}\right)$ that returns 1 when the FH link ϵ is in the feasible FH routing path set $P^m_{i,n,q}$ and 0 otherwise.

In addition, by using a set Π to collect all $\pi^m_{i,n,q}$ variables: $\Pi = \left\{\pi^m_{i,n,q}, \forall r_i \in \mathcal{R}, f_m \in \mathcal{F}, b_n \in \mathcal{B}, q \in \left[1, \left|\mathcal{P}^m_{i,n}\right|\right]\right\}$, the objective function can be rewritten in Eq. (8a) as a bilinear function in terms of both $\pi^m_{i,n,q}$ and $x_{i,j}$.

Despite the problem reformulation stated above, RRU clustering ($\mathbf{C}$) retians its binary form for several reasons. The first and foremost reason is to maintain its original definition to align it with our system model. Consider one specific example: if a single RRU is multiplexed (i.e, continuously-relaxed $c_{i,k} \in [0,1]$) by two RRU clusters, then it will be separated into two distinct sub-units, and the inference in between needs to be taken into account. However, these sub-

units violate the basis of limiting the maximum number of RRUs in one RRU cluster (cf. in Eq. (3c)) and the internal interference in one RRU is not applicable to the SINR forms of Section IV-C. Second, the computational complexity cannot be reduced even with continuously-relaxed $c_{i,k}$, because of the non-linear constraint in Eq. (3f). Therefore, to deal with binary variables in RRU clustering (i.e., $c_{i,k} = \{0, 1\}$), we utilize the combinatorial optimization approach while still satisfying the problem constraints in Eq. (3b) and Eq. (3c).

To conclude, we reformulate the problem into Eq. (8) in terms of user association ($\mathbf{X}$) and the new composite variables Π. This reformulated problem comprises the updated objective function $f(\mathbf{X}, \Pi)$ and constraints from $\mathcal{C}_1$ to $\mathcal{C}_7$. It should be noted that these updated constraints are either adopted directly from the original problem (i.e., Constraints $\mathcal{C}_1$ and $\mathcal{C}_2$ are adopted from Eq. (3d) and Eq. (3e)), modified because of the continuously-relaxed user association (i.e., Constraint $\mathcal{C}_3$ is modified from Eq. (3f)), or introduced together with new composite variable $\pi^m_{i,n,q}$ (i.e., Constraints $\mathcal{C}_4$, $\mathcal{C}_5$, and $\mathcal{C}_6$).

B. PROPOSED SOLUTION

To address the reformulated problem, a two-level turbo-based solution, as shown in Figure 3, is proposed by exploiting both combinatorial optimization and alternate convex search.

1) High-level processing

The goal of high-level processing is to exploit combinatorial optimization to update RRU clustering ($\mathbf{C}$) according to low-level outcomes. In practice, we apply the branch-and-bound method to analyze possible updates of RRU clustering that can satisfy both the constraints in Eq. (3b) and Eq. (3c), and then the one with the greatest improvement in the objective function (i.e., $f(\mathbf{X}, \Pi)$) from low-layer processing is selected to re-cluster RRUs based on $\mathbf{C}$. Afterwards, new candidates are generated from such updated RRU clustering and provided to low-level processing to analyze their respective improvement on $f(\mathbf{X}, \Pi)$. This method is terminated when there are no candidates for updating the latest RRU clustering. Thus, the final binary values of $c_{i,k}$ in $\mathbf{C}$ are determined. As shown in Figure 3, the feasible FH routing path sets $\mathcal{P}^m_{i,n}$ are also updated in high-level processing based on the user association and the composite variable values provided by low-level processing. These updated sets are provided to and utilized in low-level processing (cf. $\mathcal{P}^m_{i,n}$ in Eq. (8)).

2) Low-level processing

The goal of low-level processing is to alternatively solve our reformulated problem in Eq. (8) in terms of two aspects: (1) User association $\mathbf{X}$, and (2) joint functional split, FH routing, and BBU placement Π. As mentioned previously, both RRU clustering $\mathbf{C}$ and feasible FH routing path sets $\mathcal{P}^m_{i,n}$ are provided by high-level processing. In specific, there are three alternating stages of low-level processing. The first stage applies convex optimization to tackle a sub-problem comprising the objective function $f_\Pi(\mathbf{X})$ together with constraints from $\mathcal{C}_1$ to $\mathcal{C}_3$. It can be observed that $f_\Pi(\mathbf{X})$ is the same as $f(\mathbf{X}, \Pi)$ in Eq. (8a) but with a fixed Π. Then, in the second stage, a sub-problem including the objective function $f_\mathbf{X}(\Pi)$ (i.e., with a fixed $\mathbf{X}$) and constraints from $\mathcal{C}_4$ to $\mathcal{C}_6$ will also be tackled by convex optimization. Subsequently, the third stage rounds user association values $x_{i,j}$ into binary forms, and then updates the expected SINR accordingly, i.e., $E[\mathbf{T}|\mathbf{X}]$. Finally, these alternating stages are terminated after the objective function converges or the maximum cycle count is reached.

$$\text{maximize} \quad f(\mathbf{X}, \Pi) = \sum_{j=1}^{|\mathcal{U}|} \sum_{i=1}^{|\mathcal{R}|} \sum_{m=1}^{|\mathcal{F}|} \sum_{n=1}^{|\mathcal{B}|} \sum_{q=1}^{|\mathcal{P}^m_{i,n}|} x_{i,j} \cdot n_i \cdot \pi^m_{i,n,q} \cdot \log\left(1 + E[\tau_{j,m}|\mathbf{X}]\right) \tag{8a}$$

$$\text{subject to} \quad \mathcal{C}_1\text{:} \sum_{r_i \in \mathcal{R}} x_{i,j} \cdot n_i - 1 = 0, \quad \forall u_j \in \mathcal{U} \tag{8b}$$

$$\mathcal{C}_2\text{:} \; x_{i,j} - \overline{q}_{i,j} \leqslant 0, \quad \forall r_i \in \mathcal{R}, \; u_j \in \mathcal{U} \tag{8c}$$

$$\mathcal{C}_3\text{:} \; x_{i,j} \cdot \overline{c}_{i,k} - x_{k,j} \cdot \overline{c}_{k,i} = 0, \quad \forall r_i \neq r_k \in \mathcal{R}, \; u_j \in \mathcal{U} \tag{8d}$$

$$\mathcal{C}_4\text{:} \sum_{f_m \in \mathcal{F}} \sum_{b_n \in \mathcal{B}} \sum_{q=1}^{|\mathcal{P}^m_{i,n}|} \pi^m_{i,n,q} - 1 = 0, \quad \forall r_i \in \mathcal{R} \tag{8e}$$

$$\mathcal{C}_5\text{:} \sum_{q=1}^{|\mathcal{P}^m_{i,n}|} \pi^m_{i,n,q} \cdot \overline{c}_{i,k} - \sum_{q=1}^{|\mathcal{P}^m_{k,n}|} \pi^m_{k,n,q} \cdot \overline{c}_{k,i} = 0, \forall r_i \neq r_k \in \mathcal{R}, f_m \in \mathcal{F}, b_n \in \mathcal{B} \tag{8f}$$

$$\mathcal{C}_6\text{:} \sum_{i=1}^{|\mathcal{R}|} \sum_{n=1}^{|\mathcal{B}|} \sum_{m=1}^{|\mathcal{F}|} \sum_{q=1}^{|\mathcal{P}^m_{i,n}|} \pi^m_{i,n,q} \cdot I\left(\epsilon, P^m_{i,n,q}\right) \cdot W_R(\mathbf{X}, f_m) - l_\epsilon \leqslant 0, \forall \epsilon \in \mathcal{E} \tag{8g}$$

$$\mathcal{C}_7\text{:} \; x_{i,j}, \pi^m_{i,n,q} \in [0, 1], \forall r_i \in \mathcal{R}, x_j \in \mathcal{U}, f_m \in \mathcal{F}, b_n \in \mathcal{B}, q \in \left[1, \left|\mathcal{P}^m_{i,n}\right|\right] \tag{8h}$$

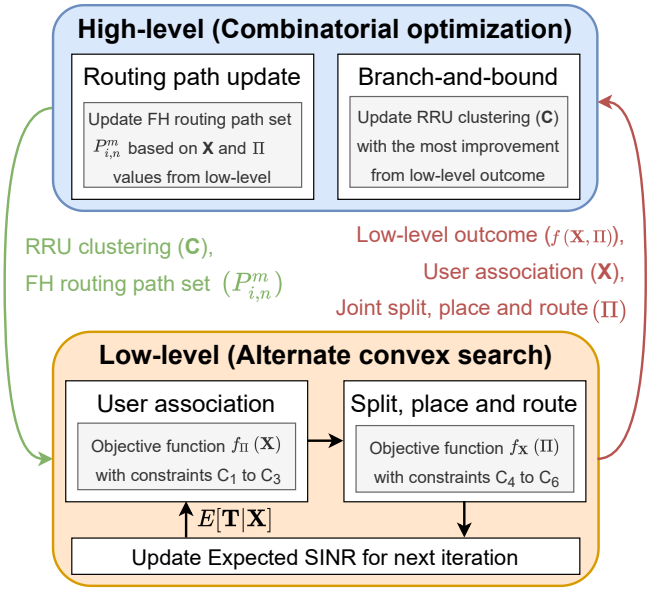


FIGURE 3: Proposed two-level turbo-based solution.

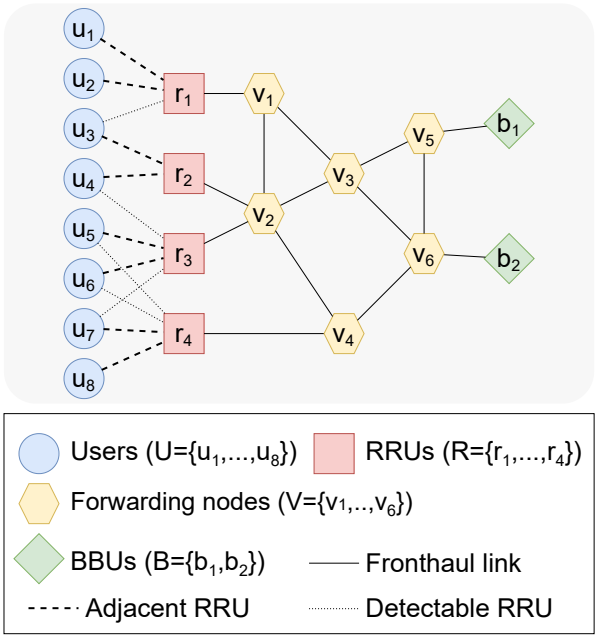


FIGURE 4: Considered small-scale network topology.

TABLE 6: Applied simulation parameters.

Parameter	Value
Channel gain variance ($\sigma^2_{i,j}$)	Adjacent RRU: 4 Detectable RRUs: 2 Other RRUs: 0.1
Minimum user association threshold (γ_{th})	1
Antenna number of each RRU (M)	4
AWGN variance (N_0)	1
ICI power ratio (γ_{ici})	3E-3
Maximum delay for reception ($T^{\max}_{rx}$)	3 ms
FH link delay ($t_\epsilon, \forall\epsilon$)	50 μs
FH link capacity ($l_\epsilon, \forall\epsilon$)	1 Gbps
Maximum iteration number for convex optimization	1000
Maximum cycle count for alternate convex search	10

FIGURE 5: Comparison between proposed and optimal solutions.

VI. PERFORMANCE EVALUATIONS

To present the performance of our proposed solution and compare it with other related works, we provide the simulation results in this section for two network topologies of different scales.

A. SMALL-SCALE NETWORK TOPOLOGY

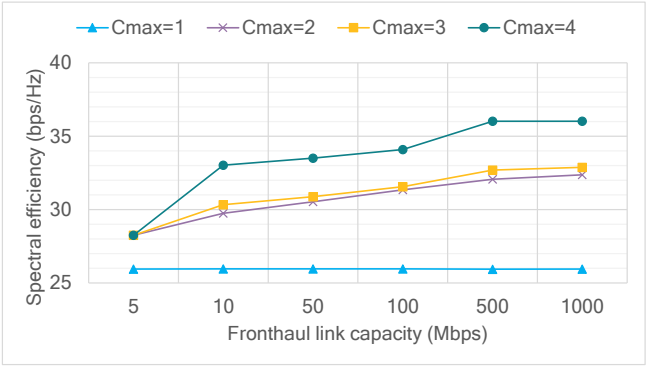
First, we consider the small-scale topology shown in Figure 4, in which the adjacent RRU (cf. Eq. (1a)) and the detectable RRUs (cf. Eq. (1b)) of each user are represented in different line styles. Also, Table 6 summarizes the simulation parameters. As mentioned in Section IV-B4, we measure the RRU processing time and the BBU processing time over the OpenAirInterface platform [11] respectively as $T_R(\cdot)$ and $T_B(\cdot)$, and apply the packetization scheme from our prior work in [6] to model $W_R(\cdot)$.

First, we compare our proposed solution to the optimal one (via exhaustive search) in Figure 5, after limiting the number of RRUs in a cluster to no larger than 3 (i.e., $C_{max} = 3$) and executing the simulation more than 10000 times (each with different random seeds). We can see that both solutions reach a similar average network spectral efficiency with no significant difference in the standard deviation (a half of error bar), whereas the execution time is reduced by a factor of 10 on average after applying the proposed solution. To give more details, two formed RRU clusters are presented in Table 7. These results match our observations in Section IV-C4 when deriving the expected SINR of different functional splits. In the first RRU cluster, split E is applied because it contains only one RRU; therefore, all functional splits provide identical performance. By contrast, the second RRU cluster exploits the joint reception CoMP scheme (splits A and B) over the three RRUs in this cluster to boost the performance of its six associated users.

Next, the network spectral efficiency is shown in Figure 6 for different FH link capacities (l_ϵ) and different maximum numbers of RRUs in an RRU cluster (C_{max}). We notice that the performance remains the same even with different FH

TABLE 7: Two formed RRU clusters in small-scale topology when l_ϵ=1 Gbps and $C_{max} = 3$.

Cluster	RRU	User	Functional split	Anchor BBU
1	$\{r_1\}$	$\{u_1, u_2\}$	1%: Split D, 99%: Split E	63%: b_1, 37%: b_2
2	$\{r_2, r_3, r_4\}$	$\{u_3, \cdots, u_8\}$	29%: Split A, 71%: Split B	38%: b_1, 62%: b_2



Network condition	Joint reception (Splits A, B, and C)	Soft symbol combination (Split D)	Transport block selection (Split E)
$C_{max} = 4$, l_ϵ=5 Mbps	Cannot form 4-RRU cluster (100% split E)		
$C_{max} = 4$, l_ϵ=10 Mbps	0.0%	1.0%	99.0%
$C_{max} = 4$, l_ϵ=100 Mbps	0.0%	36.6%	63.4%
$C_{max} = 4$, l_ϵ=1 Gbps	57.5%	42.5%	0.0%

FIGURE 6: Performance under different FH link capacity and C_{max}.

link capacities when C_{max} is 1; thus, the FH network plays little role in this condition. By contrast, an increased FH link capacity has an effect when C_{max} is greater than 1. Moreover, a better BBU hosting capability, i.e., a larger C_{max}, can also enhance the performance; however, the marginal gain of increasing C_{max} depends on the network topology. For instance, when $C_{max} = 2$, two RRU clusters are formed as $c_1 = \{r_1, r_2\}$ and $c_2 = \{r_3, r_4\}$ to serve the first four and the last four users, respectively. Because there is no strong interference between these two groups of users (cf. Figure 4); thus, little improvement is seen after increasing C_{max} to 3. When C_{max} is increased to 4, we can see from the table in Figure 6 that better CoMP schemes can be utilized to boost network spectral efficiency after enlarging the FH link capacity. Nevertheless, when the FH link capacity is small, i.e., 5 Mpbs, increasing C_{max} takes no effect.

In short, the network performance depends on several design factors, and a joint consideration can bring further improvement, e.g., 1.39-times and 1.10-times network spectral efficiency when compared the proposed solution with fixed clustering (**C**) and functional split (**F**) ones, respectively.

B. MEDIUM-SCALE NETWORK TOPOLOGY

The medium-scale topology shown in Figure 7 is then investigated, and the same parameters listed in Table 6 are applied. First, we extensively evaluate the performance of spectral efficiency in Figure 9, with varying values of the FH link delay (t_ϵ) and the maximum number of RRUs in an RRU cluster (C_{max}). It should be noted that the FH link delay does not play a key role on network performance if the BBU hosting capability is limited (i.e., $C_{max} \leq 8$). This is because when fewer RRUs can collaborate in the same BBU pool, the formation of RRU clusters is primarily driven by the user interference scenario (i.e., the adjacent and detectable RRUs of each user shown in Figure 7). In this sense, the RRU cluster is formed by some neighboring RRUs and anchored to the closest BBU, e.g., $\{r_7, \cdots, r_{14}\}$ is anchored to b_5.

By contrast, the FH link delay becomes more critical once the BBU can host more RRUs in a single RRU cluster. This can be clearly observed in Figure 9 when $C_{max} > 8$. The reason behind is because a larger FH link delay prevents the finding of feasible routing paths from all RRUs to a single anchor BBU. To be more specific, we provide details of the formed RRU cluster(s) in Table 8 and Table 9 respectively for $t_\epsilon = 100\mu$s and $t_\epsilon = 50\mu$s, under the most powerful BBU hosting capability (i.e., $C_{max} = 23$). Note that the two formed RRU clusters in Table 8 contain fewer than 23 RRUs, while a large RRU cluster can be built, as shown in Table 9. These results indicate that the FH link delay is a performance-limiting factor when C_{max} is large.

In addition, as shown in Table 9, split D is applied mostly to all RRUs. This opens up an opportunity for future performance enhancement by enlarging the FH link capacity to more than 1 Gbps, particularly for all incoming FH links to the two anchor BBUs (i.e., b_2 and b_5) of a single RRU cluster. Thus, a better CoMP scheme than the soft symbol combination (e.g., joint reception of splits A and B) can be applied to further boost spectral efficiency.

Moreover, in Figure 8, a comparison between our proposed solution and several related works in Section II-C is shown, with varying values of the FH link delay (t_ϵ) and the maximum number of RRUs in an RRU cluster (C_{max}). Specifically, a significant performance gain is provided by our proposed solution to these works in [20], [22], [23], [24], as summarized in the table below the same figure. This is because they only deal with a portion of the five design factors, under the assumption that the remaining factors will be fixed. Taking the works in [23], [24] as examples, they both treat RRU clustering (**C**) as a fixed value; thus, there is no spectral efficiency improvement, even with a small FH link delay and/or a powerful BBU hosting capability. In addition, in the worst case, these works take a similar execution time as our proposed solution, because they adopt similar combinatorial optimization methods, for example, backtracking and branch-and-bound; however, the proposed approach aims at a less-restricted disaggregated RAN deployment.

Finally, we present the multiplexing gain achieved by our proposed solutions and several related works in Figure 10. To quantify the multiplexing gain for both compute and network resources in the multi-segment FH network, we define the following two metrics:

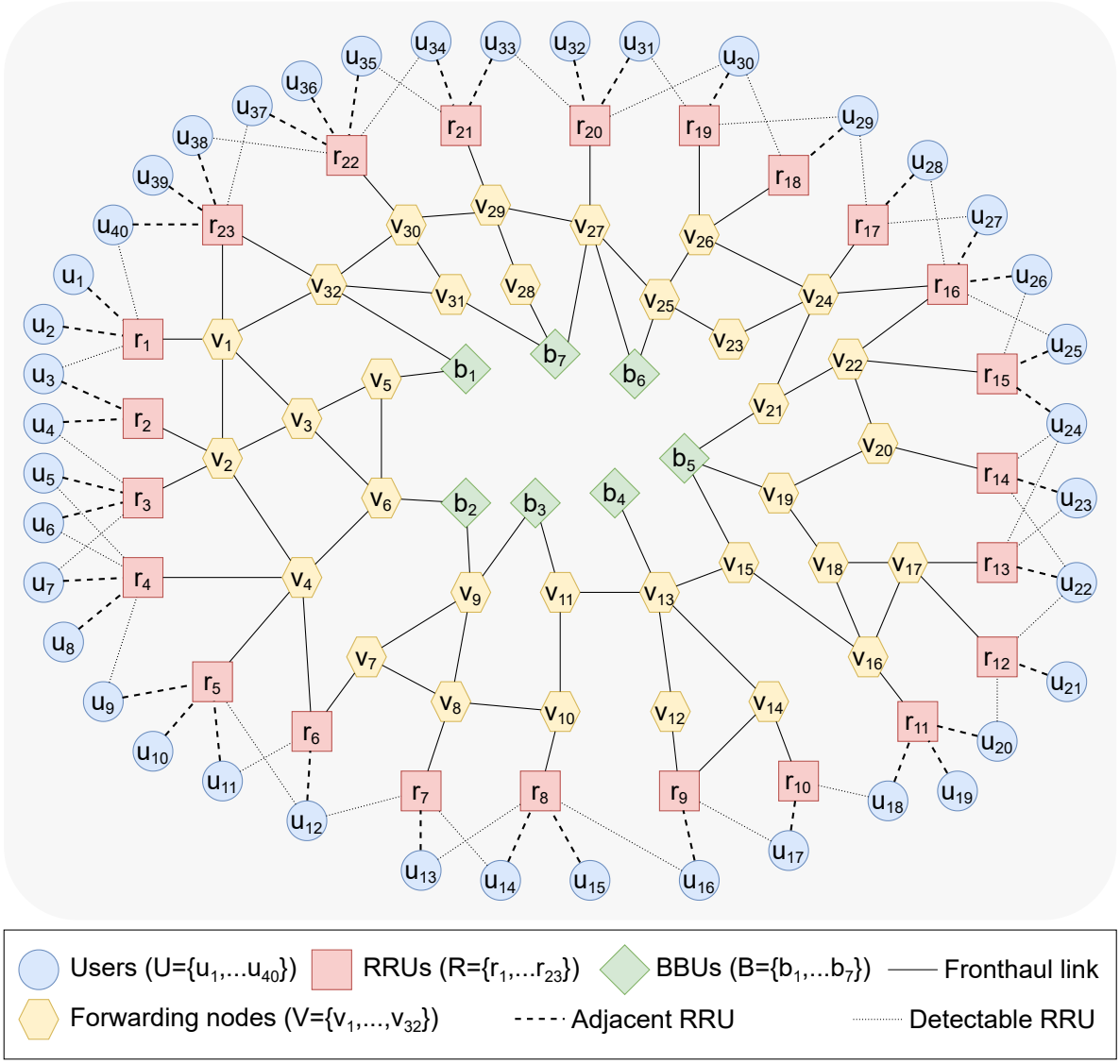


FIGURE 7: Considered medium-scale network topology.

- Compute resource multiplexing gain is defined as the ratio of the number of nodes with incoming or outgoing FH traffic to all nodes (i.e., RRUs, BBUs, and forwarding nodes).

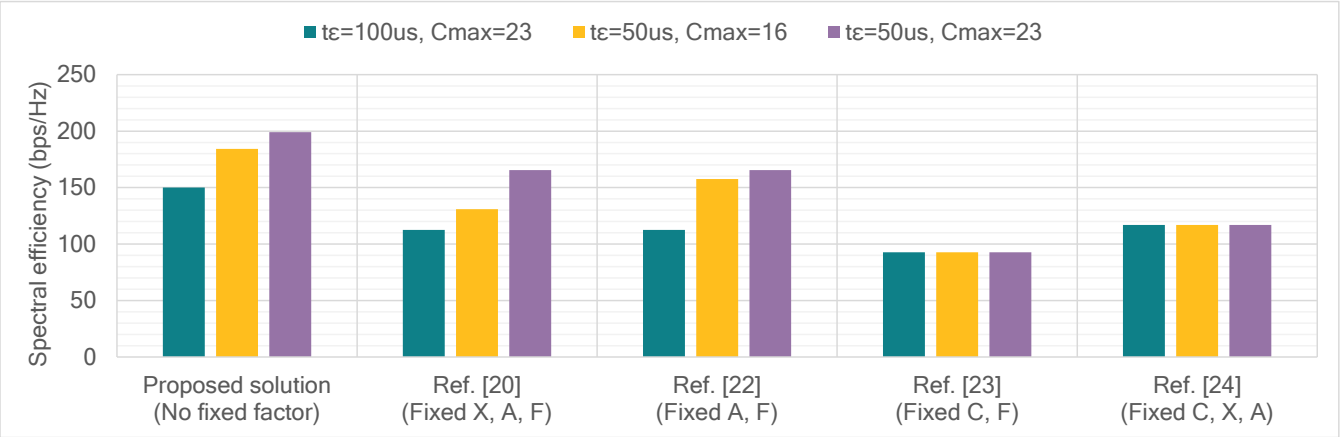
$$G_c = \frac{|\mathcal{V}|}{\sum_{v\in\mathcal{V}}\left(\sum_{\epsilon\in\delta^+(v)\cup\delta^-(v)}\sum_{r_i\in\mathcal{R}} e_{i,\epsilon} \succ 0\right)} \tag{10}$$

- Network resource multiplexing gain is defined as the ratio of the number of used FH links to the number of available FH links in the multi-segment FH network.

$$G_n = \frac{|\mathcal{E}|}{\sum_{\epsilon\in\mathcal{E}}\left(\sum_{r_i\in\mathcal{R}} e_{i,\epsilon} \succ 0\right)} \tag{11}$$

For our proposed solution, these two multiplexing gains can reach up to 1.27 (G_c) and 1.74 (G_n) among the three considered scenarios, respectively, and they are decreased when a larger RRU cluster can be formed (i.e., smaller t_ϵ or larger C_{max}). A similar trend is observed when other solutions from [20], [22], [23], [24] are applied. Moreover, we notice that our proposed solution can reach multiplexing gains similar to those in [20], [22], except for the case with the smallest FH link delay ($t_\epsilon = 50\mu s$) and the largest BBU hosting capability (cf. Table 9). The reason behind is because several extra forwarding nodes and FH links are utilized to establish a 23-RRU cluster to boost spectral efficiency, as shown in Figure 8. In contrast, our proposed solution can provide a higher multiplexing gain than those in [23], [24] because of the flexibility in forming RRU clusters (i.e., $\mathbf{C}$) in the disaggregated RAN.

In summary, full flexibility in deploying a disaggregated RAN can be obtained by dealing with all these design factors, and the results show that further performance improvement can be achieved even under the same FH network condition and BBU hosting capability.



Performance gain by proposed solution	Ref. [20]	Ref. [22]	Ref. [23]	Ref. [24]
$t_\epsilon = 100us$, $C_{max} = 23$	33.33%	33.21%	61.69%	28.31%
$t_\epsilon = 50us$, $C_{max} = 16$	40.79%	16.94%	98.66%	57.66%
$t_\epsilon = 50us$, $C_{max} = 23$	20.23%	20.23%	114.62%	70.32%

FIGURE 8: Spectral efficiency comparisons between proposed solution and several related works.

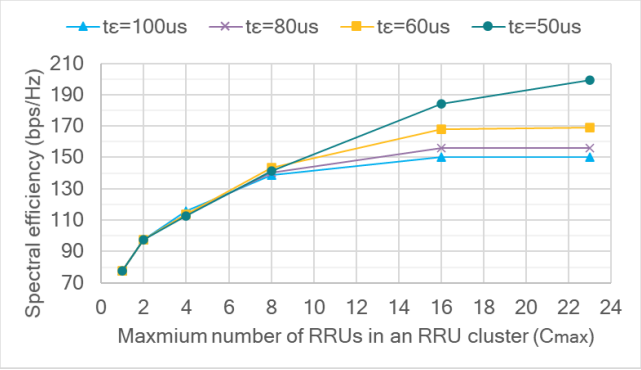


FIGURE 9: Performance under different FH link delay and C_{max}.

TABLE 8: Two formed RRU clusters in medium-scale topology when $t_\epsilon = 100\mu s$ and $C_{max} = 23$.

Cluster	RRU	User	Functional split	Anchor BBU
1	$\{r_7, \cdots, r_{14}\}$	$\{u_{13}, \cdots, u_{24}\}$	42%: Split C, 58%: Split D	100%: b_5
2	$\{r_1, \cdots, r_6, r_{15}, \cdots, r_{23}\}$	$\{u_1, \cdots, u_{12}, u_{25}, \cdots, u_{40}\}$	20%: Split B, 80%: Split D	100%: b_7

TABLE 9: One formed RRU cluster in medium-scale topology when $t_\epsilon = 50\mu s$ and $C_{max} = 23$.

Cluster	RRU	User	Functional split	Anchor BBU
1	$\{r_1, \cdots, r_{23}\}$	$\{u_1, \cdots, u_{40}\}$	2%: Split A, 4%: Split B, 94%: Split D	52%: b_2, 48%: b_5

C. DISCUSSIONS

Based on the above experiments, we can see that all five design factors should be considered for a general disaggregated RAN deployment. Specifically, their particular performance impacts occur for different BBU hosting capabilities (i.e., C_{max}). When such a hosting capability equals 1, each individual RRU is treated as one RRU cluster, and all functional splits provide identical performance. In this sense, there is no need for a large FH link capacity, either for clustering RRUs or for applying a better CoMP scheme. Thus, the FH link capacity is only scaled up in proportion to radio parameters such as radio bandwidth and antenna number. When C_{max} increases (e.g., $C_{max} \geq 2$ in Figure 4 and $2 \leq C_{max} \leq 8$ in Figure 7), RRU clusters are gradually formed from some neighboring RRUs to mitigate the interference between users. In this sense, the FH link capacity starts to play a key role in accommodating split-dependent FH traffic from several RRUs to their anchor BBU. Finally, a balance is made between forming a large RRU cluster and applying a better CoMP scheme.

In continuation to increase the BBU hosting capability (e.g., $C_{max} > 8$ in Figure 7), the FH link delay starts to be important, because it largely limits the feasible routing paths to the anchor BBU. As shown in Table 8 and Table 9, only a portion of RRUs can be clustered under a large FH link delay. Therefore, the most challenging issue is to anchor the BBU at a suitable location, where all FH traffic flows can be routed with a delay lower than the upper bound. Finally, when the FH link delay is small (e.g., $l_\epsilon \leq 50\mu s$ in Figure 7), it opens up further opportunities for applying better CoMP schemes at the cost of increasing the FH link capacity towards the anchor BBU.

As a summary, to fully exploit the benefits of RAN disaggregation, the resource evaluation of both RAN and transport network domains is necessary. Thus, the design should consider all relevant factors together - by making appropriate trade-offs between them to achieve a flexible disaggregatd RAN deployment.

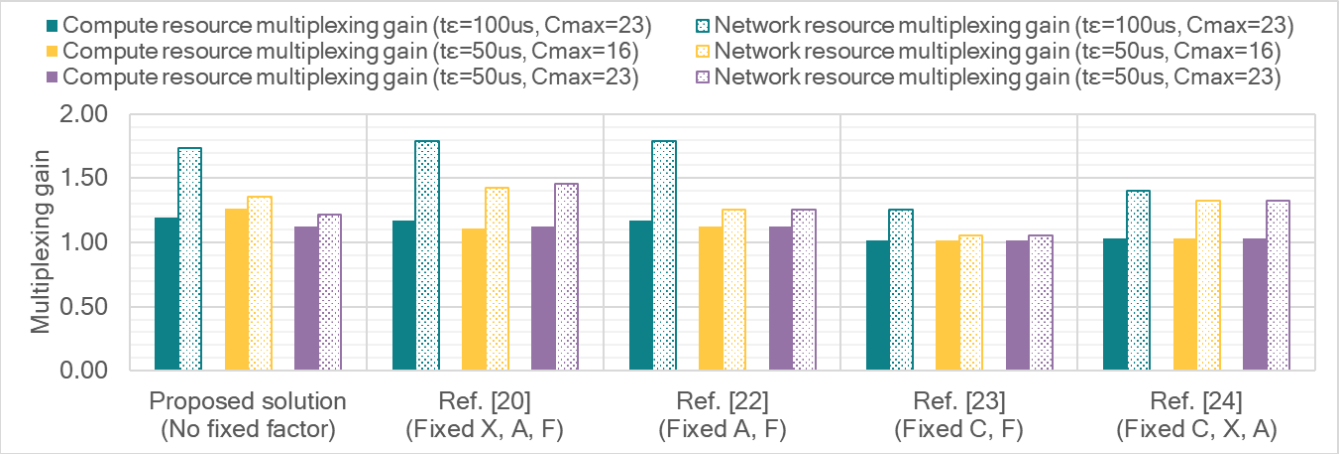


FIGURE 10: Multiplexing gain comparisons between proposed solution and several related works.

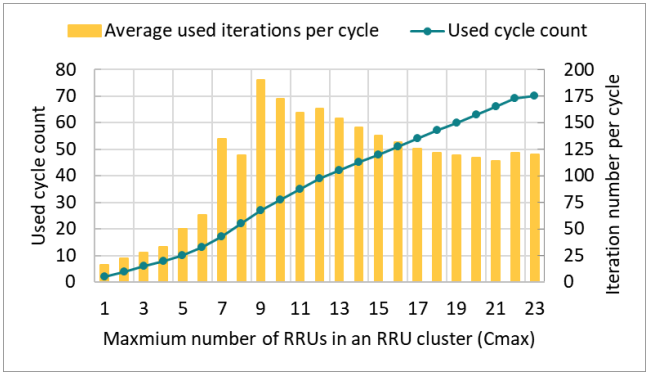


FIGURE 11: Used cycle counts and average iteration numbers per cycle under different C_{max}.

VII. EXTENSIONS AND APPLICABILITY

A. CONVERGENCE AND EXTENSIONS OF SOLUTION

Within prior experiments, the maximum cycle count is set to 10 (cf. Table 6) for the alternating stages and the maximum iteration number of convex optimization is set to 1000. These values are set based on the considered network size, the initial starting points, and the convergence characteristics for the alternating stages [36]. Regarding the initial starting point, we set the initial values of $\mathbf{X}$ and Π to associate user with their adjacent RRU and the shortest routing path with transport block selection CoMP scheme (i.e., split E) respectively. Moreover, the number of used cycles and iterations per cycle are shown in Figure 11 for different C_{max}. We can observe that the number of cycles to make a 23-RRU cluster is around 70, which means less than 10 cycles are needed to compute one candidate for adding an extra RRU[5] and the average number of iterations per cycle is less than 200.

[5] In our simulation, the number of cycles required to add one RRU into the cluster is distributed between 1 and 6.

Furthermore, our solution can interact with solutions from different domains, e.g., MEC, or other RAN controllers. Regarding the former, several related works mentioned in Section II-C aimed to provide a joint RAN/MEC solution with their respective objectives. However, our solution can provide both BBU anchoring and functional split information for each RRU cluster as inputs for scaling operations in the MEC domain. For instance, the autonomous VNF auto-scaler in [37] can deploy Deep Reinforcement Learning (DRL) agent in the MEC domain to scale the number of VNFs based on the dynamic workload information to obtain the delay target and reduce Service Level Aggregation (SLA) violations. Regarding the latter, the resource controller of the RAN can utilize both user association and RRU clustering information provided by the proposed solution to schedule radio and computing resources. Taking vrAIn in [38] as an example, it can pool all available radio resources within each RRU cluster to serve dynamic traffic requests by all associated users and apply DRL to adjust radio and computing scheduling policies.

B. APPLICABILITY TO REAL DEPLOYMENT

To be applicable to network deployment, the proposed solution must consider different real-time constraints of various design factors. The reason behind this is to align our proposed solution with the RAN Intelligent Controller (RIC) and control apps (i.e., xApps and rApps) architecture proposed by Open RAN (O-RAN) alliance [39], in which different time granularities are employed. More specifically, we can classify all five design factors into two categories (according to [40]): (1) User association and RRU clustering can be controlled by hard-real-time control apps that require a delay guarantee, while (2) Functional split, FH network routing, and BBU anchoring can be controlled by soft-real-time control apps that require an average delay guarantee within a tolerance. Therefore, separate control apps need to be developed to realize the control logic of different design factors, and their interactions and interfaces require further study.

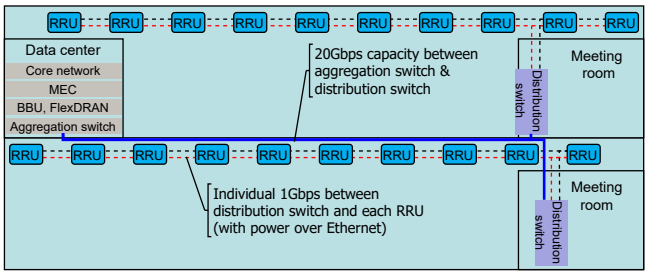


FIGURE 12: Potential indoor deployment at the EURECOM building.

Another deployment gap concerns consideration of various RAN deployment scenarios. Our proposed solution focuses on the two-tier topology (i.e., RRU and BBU) and FH network in between. The potential indoor deployment at the EURECOM building (level -3 and -4) is shown in Figure 12. However, a three-tier topology (i.e., RU, DU, and CU) must be investigated further. One main challenge we notice is the further information delay between the discolocated DU and CU, because our proposed solution is designed to be deployed at the centralized RAN entity to control a number of RUs/RRUs. In this regard, we expect that the control logic should be realized at the DU to promptly react to any fluctuation in the FH network or radio connectivity at the cost of reducing the number of RRUs in one cluster. Moreover, the user association factor should be reexamined in particular scenarios. For example, in heterogeneous deployment, user re-association with different radio characteristics (e.g., carrier frequency and radio bandwidth) needs to be avoided, because it can cause extra inter-frequency handover or synchronization procedures. In addition, re-association needs to be carefully applied to delay-critical radio bearers to avoid extra delay from re-transmission.

VIII. CONCLUSIONS

In this work, we explore the opportunity to investigate five design factors for a disaggregated RAN: (1) user association, (2) RRU clustering, (3) functional split, (4) FH network routing, and (5) BBU placement. Based on these design factors, we not only formulate an overall problem to maximize the network spectral efficiency, but also reformulate it into digestible sub-problems with new composite variables. Subsequently, a two-level turbo-based solution is provided by exploiting both combinatorial optimization and alternate convex search methods. Finally, the proposed solution is examined over two network typologies of different sizes. The numerical results show that by jointly considering five design factors, our proposed solution can achieve 1.33-times spectral efficiency compared to the state-of-the-art methods, while still provide similar multiplexing benefits (1.27 and 1.74 for compute and network resources).

On top of this work, an interesting area for future research is to inspect the real-time constraints of various design factors in real deployment, as mentioned in Sec. VII-B. Another potential area is to study the impact of performance on delay-sensitive network services by replacing the objective function and the respective constraints. Finally, joint consideration with cell-free massive Multiple-Input-Multiple-Output (MIMO) is the other possible direction for providing extra flexibility in disaggregated RAN.

.

APPENDIX A DERIVATION OF EXPECTED SINR

A. DERIVATION OF SINR FOR JOINT RECEPTION

To derive the SINR for the joint reception CoMP scheme, we introduce some additional parameter notations as follows. First, one $(M \cdot |\mathcal{R}|) \times (M \cdot |\mathcal{R}|)$ matrix is defined as $\mathbf{X}_j^{ext}$ in Eq. (13), where $x_{i,j}$ is the (i,j)-th element of user association variable matrix $\mathbf{X}$ (cf. Sec. III-B) and $\otimes$ is the Kronecker product operator. Moreover, we can remove all all-zero rows in $\mathbf{X}_j^{ext}$ and form another matrix as $\overline{\mathbf{X}}_j^{ext}$.

$$\mathbf{X}_j^{ext} = \begin{bmatrix} x_{1,j} \otimes \mathbf{I}_M & \cdots & \mathbf{0}_{M\times M} \\ \vdots & \ddots & \vdots \\ \mathbf{0}_{M\times M} & \cdots & x_{|R|,j} \otimes \mathbf{I}_M \end{bmatrix} \tag{13}$$

Afterwards, the symbol received from the j-th user to its associated RRUs in the same RRU cluster is written as $\mathbf{y}_j$ in Eq. (12a), where $\mathbf{h}_j$ is defined in Eq. (5), s_j is the j-th user transmitted symbol with zero mean and unit variance (i.e., P_s is 1, as stated in Sec. IV-C), $\boldsymbol{\eta}_j$ is the AWGN vector with zero mean and variance N_0 for all its entries, and set $\mathcal{N}_j = \{u_k : \overline{x}_{j,k} == 1\}$ contains all interfering users. In addition, we note that this formulated $\mathbf{y}_j$ is a vector of size $M_j \times 1$, where $M_j = M \cdot \left(\sum_{r_i \in \mathcal{R}} x_{i,j}\right)$ is the effective number of receiving antennas for the j-th user. For notation simplification, we define $\tilde{\mathbf{h}}_{j,k} = \overline{\mathbf{X}}_j^{ext} \cdot \mathbf{h}_k$ as the effective channel from the k-th user to all its associated RRUs. Subsequently, the MMSE receiver vector $\mathbf{w}_j$ in Eq. (12b) is applied at the anchor BBU, and the equalized symbol is written as $\hat{s}_j$ in Eq. (12c). Finally, the SINR of both splits A and B are derived as $\tau_{j,A}$ and $\tau_{j,B}$ in Eq. (6a), by expressing the variance of the first and the second terms in Eq. (12c) as numerator and denominator, respectively.

Regarding to split C, a similar derivation approach is as follows. We first write the received symbol from the j-th user to its associated RRUs in the same RRU cluster as $\mathbf{y}_{j,C}$ in Eq. (14a), in which the power of the transmitted symbol is multiplied by $0 \leq \sqrt{1 - r_{ici}} \leq 1$ and one extra noise $\mathbf{i}_j \sim CN\left(\mathbb{0}_{M_j \times 1}, \mathbf{Z}_j\right)$ is introduced because of the ICI mentioned in Sec. IV-C. It is worth noting that this extra noise for the j-th user comes from its own transmitted symbol, and the covariance matrix of this extra noise is expressed in Eq. (15), considering both the effective channel of the j-th user $\tilde{\mathbf{h}}_{j,j}$ and the ICI power ratio r_{ici}.

$$\mathbf{Z}_j = r_{ici} \cdot E\left[\tilde{\mathbf{h}}_{j,j} \cdot \tilde{\mathbf{h}}_{j,j}^H\right] \tag{15}$$

In this sense, the MMSE receiver vector is denoted as $\mathbf{w}_{j,C}$ in Eq. (14b) and the equalized symbol $\hat{s}_{j,C}$ is derived in Eq. (14c). Finally, the SINR of split C is written as $\tau_{j,C}$ in Eq. (6b), by considering the variance of the first and the

second terms in Eq. (14c) as the numerator and denominator, respectively. One can notice that, by comparing three respective sub-equations in Eq. (12) and Eq. (14), these three functional splits (splits A, B, and C) have identical received symbol, MMSE receiver vector, and equalized symbol when there is no extra ICI (i.e., r_{ici} is zero). This matches the observations in Sec. IV-C1.

B. DERIVATION OF SINR FOR SOFT SYMBOL COMBINATION

As mentioned in Sec. IV-C2, the soft symbol combination scheme first equalizes the received symbols at all distributed RRUs and then applies the combination scheme for soft symbols at the centralized BBU. We first formulate the received symbol at the i-th RRU from the j-th user as $\mathbf{y}_{i,j}$ in Eq. (16a), where s_j and $\mathcal{N}_j$ are already introduced in Appendix A-A, and $\boldsymbol{\eta}_{i,j}$ is the AWGN vector with zero mean and variance N_0 for all its elements. By applying the MMSE principle at the i-th RRU, the corresponding MMSE receiver vector is written as $\mathbf{w}_{i,j}$ in Eq. (16b), and the soft symbol $\hat{s}_{i,j}$ can be obtained from Eq. (16c). The SINR of the soft symbol $\tau^{ss}_{i,j}$ can then be formulated in Eq. (6c), by considering the variance of the first and the second terms in Eq. (16c) as the numerator and denominator, respectively. Subsequently, by using the MRC combination approach, all soft symbols are multiplied by their SINR square root, i.e., $\sqrt{\tau_{i,j}}$, and then combined at the anchor BBU as $\hat{s}_{j,D}$ in Eq. (16d). Finally, the SINR of split D $\tau_{j,D}$ can be derived in Eq. (6d) as $\hat{s}_{j,D}$.

C. DERIVATION OF EXPECTED SINR

To derive the expected SINR of splits A and B, we define a new parameter β_j in Eq. (17a), which represents the equalized channel of the j-th user after applying the MMSE principle. In specific, it is formed by multiplying the channel $\tilde{\mathbf{h}}_{j,j}$ by the MMSE receiver vector $\mathbf{w}_j^H$ and can be written in a fractional form using Sherman-Morrison formula. Then, the numerator and denominator of $\tau_{j,A}$ and $\tau_{j,B}$ in Eq. (6a) can be reformulated as Eq. (17b) as β_j^2 and $\left(\beta_j - \beta_j^2\right)$, respectively. To go one further step, we apply eigen value decomposition [41] to the interference part $\sum_{u_k \in \mathcal{N}_j} \tilde{\mathbf{h}}_{j,k} \cdot \tilde{\mathbf{h}}_{j,k}^H = \mathbf{V}_j^H \cdot \boldsymbol{\Lambda}_j \cdot \mathbf{V}_j$, in which $\boldsymbol{\Lambda}_j = diag\left(\left[\lambda_{j,1}, \cdots \lambda_{j,|\mathcal{N}_j|}\right]\right)$ is the diagonal eigenvalue matrix, and $\mathbf{V}_j$ is the unitary eigenvector matrix. Thanks to the unitary characteristic of $\mathbf{V}_j$, the eigenvector-projected channel $\check{\mathbf{h}}_j = \mathbf{V}_j \cdot \tilde{\mathbf{h}}_{j,j}$ can preserve the same characteristic as of the original channel $\tilde{\mathbf{h}}_{j,j}$ in terms of $\check{\mathbf{h}}_j^H \cdot \check{\mathbf{h}}_j = \tilde{\mathbf{h}}_{j,j}^H \cdot \tilde{\mathbf{h}}_{j,j}$. Finally, the recomposed SINR of splits A and B is presented in Eq. (17b) and it includes two parts. The first part has $|\mathcal{N}_j|$ items that are the eigenvector-projected channel power $\left|\check{h}_{j,k}\right|^2$ weighted respectively by the inverse of the interference-plus-noise power, i.e., $\frac{1}{\lambda_{j,k}+N_0}$. The second part contains $M_j - |\mathcal{N}_j|$ items with the same eigenvector-projected channel power but only weighted by the inverse of the noise power. Such intermediate results will be further utilized to derive the expected SINR.

Based on the intermediate results in Eq. (17b), we must further understand the per-entry variance in the eigenvector-projected channel $\check{\sigma}^2_{j,k}, \forall k \in [1, M_j]$ in Eq. (17c). According to its definition in the previous paragraph, we can see that such variance is contributed by both the interference eigenvector matrix (i.e., $\mathbf{V}_j$) and the original effective channel

$$\mathbf{y}_j = \overline{\mathbf{X}}_j^{ext} \cdot \mathbf{h}_j \cdot s_j + \sum_{u_k \in \mathcal{N}_j} \overline{\mathbf{X}}_j^{ext} \cdot \mathbf{h}_k \cdot s_k + \boldsymbol{\eta}_j = \tilde{\mathbf{h}}_{j,j} \cdot s_j + \sum_{u_k \in \mathcal{N}_j} \tilde{\mathbf{h}}_{j,k} \cdot s_k + \boldsymbol{\eta}_j \tag{12a}$$

$$\mathbf{w}_j = \left(\tilde{\mathbf{h}}_{j,j} \cdot \tilde{\mathbf{h}}_{j,j}^H + \sum_{u_k \in \mathcal{N}_j} \tilde{\mathbf{h}}_{j,k} \cdot \tilde{\mathbf{h}}_{j,k}^H + N_0 \cdot \mathbf{I}_{M_j} \right)^{-1} \cdot \tilde{\mathbf{h}}_{j,j} \tag{12b}$$

$$\hat{s}_j = \mathbf{w}_j^H \cdot \mathbf{y}_j = \mathbf{w}_j^H \cdot \tilde{\mathbf{h}}_{j,j} \cdot s_j + \mathbf{w}_j^H \cdot \left(\sum_{u_k \in \mathcal{N}_j} \tilde{\mathbf{h}}_{j,k} \cdot s_k + \boldsymbol{\eta}_j \right) \tag{12c}$$

$$\mathbf{y}_{j,C} = \sqrt{1 - r_{ici}} \cdot \tilde{\mathbf{h}}_{j,j} \cdot s_j + \sum_{u_k \in \mathcal{N}_j} \tilde{\mathbf{h}}_{j,k} \cdot s_k + \mathbf{i}_j + \boldsymbol{\eta}_j \tag{14a}$$

$$\mathbf{w}_{j,C} = \left(\frac{(1 - r_{ici}) \cdot \tilde{\mathbf{h}}_{j,j} \cdot \tilde{\mathbf{h}}_{j,j}^H + \sum_{u_k \in \mathcal{N}_j} \tilde{\mathbf{h}}_{j,k} \cdot \tilde{\mathbf{h}}_{j,k}^H + r_{ici} \cdot E\left[\tilde{\mathbf{h}}_{j,j} \cdot \tilde{\mathbf{h}}_{j,j}^H \right] + N_0 \cdot \mathbf{I}_{M_j}}{\sqrt{1 - r_{ici}}} \right)^{-1} \cdot \tilde{\mathbf{h}}_{j,j} \tag{14b}$$

$$\hat{s}_{j,C} = \mathbf{w}_{j,C}^H \cdot \mathbf{y}_{j,C} = \mathbf{w}_{j,C}^H \cdot \sqrt{1 - r_{ici}} \cdot \tilde{\mathbf{h}}_{j,j} \cdot s_j + \mathbf{w}_{j,C}^H \cdot \left(\sum_{u_k \in \mathcal{N}_j} \tilde{\mathbf{h}}_{j,k} \cdot s_k + \mathbf{i}_j + \boldsymbol{\eta}_j \right) \tag{14c}$$

($\tilde{\mathbf{h}}_{j,j}$). In specific, the first $|\mathcal{N}_j|$ items are affected by interfering users, i.e., the k-th entry $u_{k'}$ within the set $\mathcal{N}_j$, and thus we can write it as the interference-normalized channel power in Eq. (17c). The last $M_j - |\mathcal{N}_j|$ items will equally share all remaining channel power, and thus we can formulate it by deducting the first $|\mathcal{N}_j|$ items from the overall channel power and then dividing it by the value $M_j - |\mathcal{N}_j|$, as shown in Eq. (17c). Finally, we formulate the expected SINR of splits A and B in Eq. (7a).

Moreover, the SINR of split C can be derived in Eq. (7b) by exploiting similar steps, starting from $\tau_{j,C}$ in Eq. (6b). However, we note that there are two differences between Eq. (7a) and Eq. (7b). First, the extra $z_{j,k}$ is the k-th diagonal element of $\mathbf{Z}_j$ defined in Eq. (15) to represent the impact of extra noise $\mathbf{i}_j$ in Eq. (14a). Second, the eigenvector-projected channel $\check{\sigma}^2_{j,k}$ is multiplied by $0 \leq (1-\gamma_{ici}) \leq 1$, which can also be observed in Eq. (14a).

Additionally, to derive the expected SINR of split D and split E, we follow similar steps and recompose the soft symbol SINR at the i-th RRU from the j-th user in Eq. (18). However, two differences are observed between Eq. (17b) and Eq. (18). First, eigen value decomposition is applied to $\sum_{u_k \in \mathcal{N}_j} \mathbf{h}_{i,k} \cdot \mathbf{h}^H_{i,k} = \mathbf{V}^H_{i,j} \cdot \mathbf{\Lambda}_{i,j} \cdot \mathbf{V}_{i,j}$, and therefore we denote its k-th eigenvalues as $\lambda^{ss}_{i,j,k}$, i.e., $\mathbf{\Lambda}_{i,j} = diag\left(\left[\lambda^{ss}_{j,1}, \cdots \lambda^{ss}_{j,|\mathcal{N}_j|}\right]\right)$. Second, because all RRUs in the same RRU cluster equalize their received symbols individually, the original M_j in Eq. (17b) is replaced with M in Eq. (18). Thanks to the individual symbol equalization performed by each RRU, the interference eigenvector matrix $\mathbf{V}_{i,j}$ will have no impact on the channel variance and therefore $\check{\sigma}^2_{j,k} = \sigma^2_{j,k}$. Subsequently, the expected SINR of the soft symbols transmitted from the j-th user to the i-th RRU is expressed as $E\left[\tau^{ss}_{i,j}\right]$ in Eq. (7c). Based on the expected SINR of the soft symbols, the final expected SINR of split D and split E can be respectively formulated in Eq. (7d) and Eq. (7e), by considering their distinct CoMP schemes on soft sysymbols, i.e., combination or selection.

$$\tau^{ss}_{i,j} = \sum_{k=1}^{|\mathcal{N}_j|} \frac{1}{\lambda^{ss}_{i,j,k} + N_0} \cdot \left|\tilde{h}_{j,k}\right|^2 + \sum_{k=|\mathcal{N}_j|+1}^{M} \frac{1}{N_0} \cdot \left|\tilde{h}_{j,k}\right|^2 \tag{18}$$

$$\mathbf{y}_{i,j} = \mathbf{h}_{i,j} \cdot s_j + \sum_{u_k \in \mathcal{N}_j} \mathbf{h}_{i,k} \cdot s_k + \boldsymbol{\eta}_{i,j} \tag{16a}$$

$$\mathbf{w}_{i,j} = \left(\mathbf{h}_{i,j} \cdot \mathbf{h}^H_{i,j} + \sum_{u_k \in \mathcal{N}_j} \mathbf{h}_{i,k} \cdot \mathbf{h}^H_{i,k} + N_0 \cdot \mathbf{I}_M\right)^{-1} \cdot \mathbf{h}_{i,j} \tag{16b}$$

$$\hat{s}_{i,j} = \mathbf{w}^H_{i,j} \cdot \mathbf{y}_{i,j} = \mathbf{w}^H_{i,j} \cdot \mathbf{h}_{i,j} \cdot s_j + \mathbf{w}^H_{i,j} \cdot \left(\sum_{u_k \in \mathcal{N}_j} \mathbf{h}_{i,k} \cdot s_k + \boldsymbol{\eta}_{i,j}\right) \tag{16c}$$

$$\hat{s}_{j,D} = \sum_{r_i \in \Gamma_j} \sqrt{\tau_{i,j}} \cdot \hat{s}_{i,j} = \sum_{r_i \in \Gamma_j} \sqrt{\tau_{i,j}} \cdot \mathbf{w}^H_{i,j} \cdot \mathbf{h}_{i,j} \cdot s_j + \sum_{r_i \in \Gamma_j} \sqrt{\tau_{i,j}} \cdot \mathbf{w}^H_{i,j} \cdot \left(\sum_{u_k \in \mathcal{N}_j} \mathbf{h}_{i,k} \cdot s_k + \boldsymbol{\eta}_{i,j}\right) \tag{16d}$$

$$\beta_j = \mathbf{w}^H_j \cdot \tilde{\mathbf{h}}_{j,j} = \frac{\tilde{\mathbf{h}}^H_{j,j} \cdot \left(\sum_{u_k \in \mathcal{N}_j} \tilde{\mathbf{h}}_{j,k} \cdot \tilde{\mathbf{h}}^H_{j,k} + N_0 \cdot \mathbf{I}_{M_j}\right)^{-1} \cdot \tilde{\mathbf{h}}_{j,j}}{1 + \tilde{\mathbf{h}}^H_{j,j} \cdot \left(\sum_{u_k \in \mathcal{N}_j} \tilde{\mathbf{h}}_{j,k} \cdot \tilde{\mathbf{h}}^H_{j,k} + N_0 \cdot \mathbf{I}_{M_j}\right)^{-1} \cdot \tilde{\mathbf{h}}_{j,j}} \tag{17a}$$

$$\begin{aligned} \tau_{j,A} = \tau_{j,B} &= \frac{(\beta_j)^2}{\beta_j - (\beta_j)^2} = \tilde{\mathbf{h}}^H_{j,j} \cdot \left(\sum_{u_k \in \mathcal{N}_j} \tilde{\mathbf{h}}_{j,k} \cdot \tilde{\mathbf{h}}^H_{j,k} + N_0 \cdot \mathbf{I}_{M_j}\right)^{-1} \cdot \tilde{\mathbf{h}}_{j,j} \\ &= \tilde{\mathbf{h}}^H_{j,j} \cdot \left(\mathbf{V}^H_j \cdot \mathbf{\Lambda}_j \cdot \mathbf{V}_j + N_0 \cdot \mathbf{I}_{M_j}\right)^{-1} \cdot \tilde{\mathbf{h}}_{j,j} = \check{\mathbf{h}}^H_j \cdot \left(\mathbf{\Lambda}_j + N_0 \cdot \mathbf{I}_{M_j}\right)^{-1} \cdot \check{\mathbf{h}}_j \\ &= \sum_{k=1}^{|\mathcal{N}_j|} \frac{1}{\lambda_{j,k} + N_0} \cdot \left|\check{h}_{j,k}\right|^2 + \sum_{k=|\mathcal{N}_j|+1}^{M_j} \frac{1}{N_0} \cdot \left|\check{h}_{j,k}\right|^2 \end{aligned} \tag{17b}$$

$$\check{\sigma}^2_{j,k} = E\left[\left|\check{h}_{j,k}\right|^2\right] = \begin{cases} \frac{\sum_{r_i \in \Gamma_j} \sigma^2_{i,j} \cdot \sigma^2_{i,k'}}{\sum_{r_i \in \Gamma_j} \sigma^2_{i,k'}}, & 1 \leq k \leq |\mathcal{N}_j|, u_{k'} \text{ is } k\text{-th entry in } \mathcal{N}_j \\ \frac{M \cdot \sum_{r_i \in \Gamma_j} \sigma^2_{i,j} - \sum_{k=1}^{|\mathcal{N}_j|} \check{\sigma}^2_{j,k}}{M_j - |\mathcal{N}_j|}, & |\mathcal{N}_j| < k \leq M_j \end{cases} \tag{17c}$$

REFERENCES

[1] P. Rost, C. J. Bernardos, A. De Domenico, M. Di Girolamo, M. Lalam, A. Maeder, D. Sabella, and D. Wübben, "Cloud technologies for flexible 5G radio access networks," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 68–76, May 2014.

[2] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for mobile networks - A technology overview," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 405–426, 1st Quart. 2015.

[3] A. de la Oliva, J. A. Hernandez, D. Larrabeiti, and A. Azcorra, "An overview of the CPRI specification and its application to C-RAN-based LTE scenarios," *IEEE Commun. Mag.*, vol. 54, no. 2, pp. 152–159, Feb. 2016.

[4] F. Qamar, K. B. Dimyati, M. N. Hindia, K. A. B. Noordin, and A. M. Al-Samman, "A comprehensive review on coordinated multi-point operation for LTE-A," *Comput. Netw.*, vol. 123, pp. 19–37, Aug. 2017.

[5] J. Liu, S. Zhou, J. Gong, Z. Niu, and S. Xu, "Statistical multiplexing gain analysis of heterogeneous virtual base station pools in cloud radio access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 8, pp. 5681–5694, Aug. 2016.

[6] C.-Y. Chang, R. Schiavi, N. Nikaein, T. Spyropoulos, and C. Bonnet, "Impact of packetization and functional split on C-RAN fronthaul performance," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–7.

[7] D. Wubben, P. Rost, J. S. Bartelt, M. Lalam, V. Savin, M. Gorgoglione, A. Dekorsy, and G. Fettweis, "Benefits and impact of cloud computing on 5G signal processing: Flexible centralization through Cloud-RAN," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 35–44, Nov. 2014.

[8] *Task Force: Standard for Radio over Ethernet encapsulations and mapping*, IEEE Standards 1904.3, 2015.

[9] C.-Y. Chang, N. Nikaein, R. Knopp, T. Spyropoulos, and S. S. Kumar, "FlexCRAN: A flexible functional split framework over Ethernet fronthaul in Cloud-RAN," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, France, May 2017, pp. 1–7.

[10] K. Sundaresan, M. Y. Arslan, S. Singh, S. Rangarajan, and S. V. Krishnamurthy, "Fluidnet: A flexible cloud-based radio access network for small cells," *IEEE/ACM Trans. Netw.*, vol. 24, no. 2, pp. 915–928, Apr. 2016.

[11] N. Nikaein, "Processing radio access network functions in the cloud: Critical issues and modeling," in *Proc. ACM 6th Int. Workshop Mobile Cloud Comput. Services*, Sep. 2015, pp. 36–43.

[12] *Small cell virtualization functional splits and use cases*, SCF, Beijing, China, 2015.

[13] C.-L. I, Y. Yuan, J. Huang, S. Ma, C. Cui, and R. Duan, "Rethink fronthaul for soft RAN," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 82–88, Sep. 2015.

[14] *Further study on critical C-RAN technologies v1.0*, NGMN Alliance, Frankfurt, Germany, 2015.

[15] *Study on new radio access technology: Radio access architecture and interfaces (Release 14)*, 3GPP Standard TR 38.801, Mar. 2017.

[16] *Common Public Radio Interface: eCPRI Interface Specification (V1.0)*, CPRI, Bangalore, India, Aug. 2017.

[17] 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.

[18] R. M. Rao, M. Fontaine, and R. Veisllari, "A reconfigurable architecture for packet based 5G transport networks," in *Proc. IEEE 5G World Forum (5GWF)*. IEEE, Jul. 2018, pp. 474–477.

[19] H. Zhang, H. Liu, C. Jiang, X. Chu, A. Nallanathan, and X. Wen, "A practical semi-dynamic clustering scheme using affinity propagation in cooperative picocells," *IEEE Trans. Veh. Technol.*, vol. 64, no. 9, pp. 4372–4377, Sep. 2015.

[20] K. Boulos, M. El Helou, and S. Lahoud, "RRH clustering in cloud radio access networks," in *Proc. IEEE ICAR*, Oct. 2015, pp. 1–6.

[21] M. M. U. Rahman, H. Ghauch, S. Imtiaz, and J. Gross, "RRH clustering and transmit precoding for interference-limited 5G CRAN downlink," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2015, pp. 1–7.

[22] H. Taleb, M. E. Helou, K. Khawam, S. Lahoud, and S. Martin, "Joint user association and RRH clustering in cloud radio access networks," in *Proc. 10th Int. Conf. Ubiquitous Future Netw. (ICUFN)*, Jul. 2018, pp. 376–381.

[23] J. Yao and N. Ansari, "QoS-aware joint BBU-RRH mapping and user association in Cloud-RANs," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 4, pp. 881–889, Dec. 2018.

[24] 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.

[25] Y. Xiao, J. Zhang, and Y. Ji, "Can fine-grained functional split benefit to the converged optical-wireless access networks in 5G and beyond?" *IEEE Trans. Netw. Service Manage.*, vol. 17, no. 3, pp. 1774–1787, 2020.

[26] I. Koutsopoulos, "The impact of baseband functional splits on resource allocation in 5G radio access networks," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, 2021, pp. 1–10.

[27] P. Shantharama, A. S. Thyagaturu, N. Karakoc, L. Ferrari, M. Reisslein, and A. Scaglione, "LayBack: SDN management of multi-access edge computing (MEC) for network access services and radio resource sharing," *IEEE Access*, vol. 6, pp. 57 545–57 561, 2018.

[28] A. Garcia-Saavedra, X. Costa-Perez, D. J. Leith, and G. Iosifidis, "FluidRan: Optimized vRAN/MEC orchestration," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, 2018, pp. 2366–2374.

[29] M. Sun, L. Pu, J. Zhang, and J. Xu, "Matryoshka: Joint resource scheduling for cost-efficient MEC in NGFI-based C-RAN," in *Proc. IEEE Int. Conf. Commun. (ICC)*. IEEE, 2019, pp. 1–7.

[30] A. Wilzeck, Q. Cai, M. Schiewer, and T. Kaiser, "Effect of multiple carrier frequency offsets in MIMO SC-FDMA systems," in *Proc. Int. ITG/IEEE Workshop Smart Antennas*, Feb. 2007, pp. 1–7.

[31] M. Rupp, S. Schwarz, and M. Taranetz, The Vienna LTE-Advanced Simulators. Singapore, Singapore: Springer, 2016.

[32] A. Goldsmith, Wireless communications. Cambridge, U.K.: Cambridge university press, 2005.

[33] A. M. Tulino and S. Verdú, Random Matrix Theory and Wireless Communications. Norwell, MA, USA: Now Publishers Inc, 2004.

[34] H. Kellerer, U. Pferschy, and D. Pisinger, Knapsack problems. Cham, Switzerland: Springer, 2004.

[35] J. Gorski, F. Pfeuffer, and K. Klamroth, "Biconvex sets and optimization with biconvex functions: A survey and extensions," *Math. Methods Oper. Res.*, vol. 66, no. 3, pp. 373–407, Dec. 2007.

[36] S. Boyd, S. P. Boyd, and L. Vandenberghe, Convex optimization. Cambridge, U.K.: Cambridge university press, 2004.

[37] P. Soto, D. De Vleeschauwer, M. Camelo, Y. De Bock, K. De Schepper, C.-Y. Chang, P. Hellinckx, J. F. Botero, and S. Latré, "Towards autonomous VNF auto-scaling using deep reinforcement learning," in *Proc. 8th Int. Conf. Softw. Defined Syst. (SDS)*, Dec. 2021, pp. 162–167.

[38] J. A. Ayala-Romero, A. Garcia-Saavedra, M. Gramaglia, X. Costa-Perez, A. Banchs, and J. J. Alcaraz, "VrAIn: A deep learning approach tailoring computing and radio resources in virtualized RANs," in *Proc. 25th Annu. Int. Conf. Mobile Comput. Netw*, 2019, pp. 1–16.

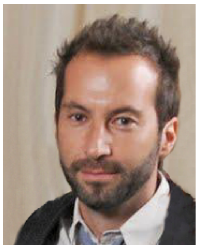
[39] O-RAN Minimum Viable Plan and Acceleration towards Commercialization, O-RAN alliance, Alfter Germany, Jun. 2021.

[40] C.-Y. Chang and N. Nikaein, "Closing in on 5G control apps: enabling multiservice programmability in a disaggregated radio access network," *IEEE Veh. Technol. Mag.*, vol. 13, no. 4, pp. 80–93, 2018.

[41] N. Kim, Y. Lee, and H. Park, "Performance analysis of MIMO system with linear MMSE receiver," *IEEE Trans. Wireless Commun.*, vol. 7, no. 11, pp. 4474–4478, Nov. 2008.

CHIA-YU CHANG (Member, IEEE) received the Ph.D. degree from Sorbonne Université, France, in 2018. He is currently a Network System Researcher at Nokia Bell Labs. He has more than 12 years of experience in algorithm/protocol research on communication systems. He has extensive research experience at both academic and industrial laboratories, including EURECOM Research Institute, MediaTek, Huawei Swedish Research Center, and Nokia Bell Labs. He has participated in several European Union's collaborative research and innovation projects, such as COHERENT, SLICENET, 5G-PICTURE, 5Growth, and DAEMON. His research interests include wireless communication, computer networking, edge computing, with particular interest in network slicing, traffic engineering, and AI/ML supported network control.

NAVID NIKAEIN received the Ph.D. degree in communication systems from the Swiss Federal Institute of Technology EPFL, in 2003. He is currently a Professor with the Communication System Department, EURECOM, leading a group focusing on experimental system research in the area of agile and flexible wireless networking and computing systems. His research interests include the areas of wireless access layer techniques, open RAN and CN architectures and interfaces, data-driven wireless computing, wireless system prototyping, and emulation/simulation platforms. He is a Board Member of OpenAirInterface.org. software alliance and leading the Mosaic5G project group whose goal is to provide software-based 4G/5G service delivery platforms.



THRASYVOULOS SPYROPOULOS received the Diploma degree in electrical and computer engineering from the National Technical University of Athens, Greece, and the Ph.D. degree in electrical engineering from the University of Southern California. He was a Postdoctoral Researcher at Inria, and then, a Senior Researcher with the Swiss Federal Institute of Technology (ETH) Zurich. He is currently an Assistant Professor at EURECOM, Sophia Antipolis. He was a recipient of the Best Paper Award from IEEE SECON, in 2008, and the IEEE WoWMoM, in 2012.

KOEN DE SCHEPPER received the M.Sc. degree in industrial sciences (electronics software engineering) from IHAM Antwerpen, Belgium. He joined Nokia (then Alcatel), in 1990, where during the first 18 years, he was the Platform Development Leader and a Systems Architect. He has been with Bell Labs for the past 13 years, currently in the Network Systems and Security Research Laboratory. Before, he worked mainly on transport layer protocols (L4) that support scalable (SCAP) and low latency (L4S) content delivery. His current research interests include programmable data plane traffic management, customizable network slicing, and AI supported dynamic service and network control.