Joint Radio Resource Allocation and Beamforming Optimization for Industrial IoT in SDN-based Virtual Fog-RAN 5G-and-Beyond Wireless Environments

Payam Rahimi, Chrysostomos Chrysostomou, Haris Pervaiz, Vasos Vassiliou and Qiang Ni

***Abstract*—Fog computing based radio access network (Fog-RAN) leveraging the software-defined networking (SDN) and network function virtualization (NFV) is the most promising solution to offer real-time support for the massive number of connected devices in the industrial internet of things (IIoT) networks. However, designing an optimal dynamic radio resource allocation to handle the fluctuating traffic loads is critical. In this paper, a novel architectural design of an SDN based virtual Fog-RAN is proposed, in which we jointly study radio resource allocation and transmit beamforming to improve resource utilization and IIoT users' satisfaction, by minimizing the network power consumption (NPC) and maximizing the achievable sum-rate (ASR), simultaneously. To this end, we first formulate a mixed-integer nonlinear problem (MINLP) to optimize the physical resource block (PRB) allocation, the assignment of user equipments (UEs) and radio unit (RU), and the downlink transmit beamforming, by considering imperfect channel state information (CSI). To solve the intractable MINLP, we exploit the successive convex approximation (SCA) approach. Then, we formulate a multiple knapsack problem (MKP) to optimize the assignment between RUs and virtual baseband units (vBBUs), by exploiting the set of active RUs minimized in the previous problem. We solve the formulated MKP by decomposing the dual problems and solving them through the dual descent (DD) method. Through performance analysis, we show the proposed approach provides a high users' satisfaction rate, maximizes the ASR and minimizes the NPC, and provides better savings, in terms of the number of radio and baseband resources utilized, than its counterparts.**

***Index Terms*—Radio resource allocation, Beamforming, Virtual fog computing based radio access network, Industrial internet of things, Network function virtualization, Software-defined networking.**

I. Introduction

With the rapid growth in the number of connected devices, data traffic is increasing enormously in the Internet of Things (IoT) networks. It is projected that future IoT networks have to enhance the network capacity 1000-fold in the following decade to meet the ever-increasingly traffic demand [1]. To address this issue, integration of industrial IoT (IIoT) in 5G and beyond networks is a key solution [2].

The cloud computing based radio access network (C-RAN) is a potential architecture to cope with the massive number of IIoT devices by applying cloud computing for handling the huge traffic loads [3]. In conventional C-RAN architecture, the baseband units (BBUs) are decomposed from base stations (BSs) and aggregated into the BBU pool at the cloud for providing the centralized signal processing, whereas the radio units (RUs) only deal with transferring and receiving signals. Through this way, the RUs are effectively installed closer to IIoT devices with an affordable operational cost [4]. However, C-RAN performance strictly relies on the fronthaul capacity [5].

Taking the advantages of both C-RAN and fog computing [6] that expand the cloud resources and services to the network edge, fog computing-based RAN (Fog-RAN) has been proposed to alleviate the challenges of the conventional C-RAN architecture by migrating some BBU processing functions to the edge of the network [6]. This allows the IIoT devices to utilize the BBU resources at the edge nodes, which dramatically increases the traffic delivery rate and energy-efficiency [7]. Fog-RAN can be deployed centralized by exploiting the software-defined networking (SDN) technology and network function virtualization (NFV). The SDN technology offers significant flexibility, abstraction, and programmability for the network communication by decoupling the control plane and data plane [8]. Moreover, the NFV orchestrator located in the SDN controller instantiates the virtual BBUs (vBBUs) on-demand to flexibly manage the BBU resources [9]. Through this way, Fog-RAN offers better adaptation for the dynamic traffic and radio environment.

However, Fog-RAN requires a dynamic radio resource allocation to support the massive number of IIoT devices due to the huge traffic loads and the environment dynamicity. Minimizing the required vBBUs and RUs to handle the traffic loads significantly improves network capacity and energy efficiency [10]. To realize the dynamic resource allocation, RUs should be allocated optimally to user equipments (UEs) and selectively switched off/on based on the traffic loads, which dramatically reduces the power consumption and overhead in the fronthaul transport network. Moreover, vBBUs should be dynamically instantiated on-demand due to the shortage of computational resources [5]. Furthermore, it is critical to improve radio resource utilization by increasing the number of UEs receiving high-quality service [11]. To meet these objectives, the joint optimization of the radio resource allocation and the transmit beamforming [12] is pivotal to design an optimal dynamic radio resource allocation.

This paper proposes a novel architectural design of an SDN-based virtual Fog-RAN in 5G-and-beyond wireless environments to significantly improve the radio resource utilization and the IIoT users' satisfaction, exploiting the flexibility offered by SDN and NFV to manage the radio and baseband resources, and the processing of the baseband signals by fog computing at the network edge.

To meet these objectives, we consider the downlink frequency division duplex (FDD) transmission in the proposed architecture, where we investigate the joint radio resource allocation and the transmit beamforming optimization problem to minimize the network power consumption (NPC) and max-

Payam Rahimi and Chrysostomos Chrysostomou (corresponding author) are with the Department of Electrical Engineering, Computer Engineering and Informatics, Frederick University, Nicosia, Cyprus (e-mail: payam.rahimi@stud.frederick.ac.cy; ch.chrysostomou@frederick.ac.cy).

Haris Pervaiz and Qiang Ni are with the School of Computing and Communications, Lancaster University, Lancaster, UK (e-mail: h.b.pervaiz@lancaster.ac.uk; q.Ni@lancaster.ac.uk).

Vasos Vassiliou is with the Department of Computer Science, University of Cyprus and CYENS Research Center, Nicosia, Cyprus (e-mail: vasosv@cs.ucy.ac.cy).

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 739578 and the Government of the Republic of Cyprus through the Deputy Ministry of Research, Innovation and Digital Policy.

imize the achievable sum-rate (ASR) by optimizing the physical resource block (PRB) allocation, the UE-RU assignment, and the downlink transmit beamforming. Then, by exploiting the minimized number of active RUs, we optimize the RU-vBBU assignment.
The main contributions of this work are summarised as:

- A novel architectural design to utilize the SDN and NFV principles for a virtual Fog-RAN in 5G-and-beyond wireless environment. In this architecture, fog computing and NFV together provide the cloud computing functionalities at the network edge. The SDN controller is in charge of controlling the RAN functionalities, in which the NFV orchestrator deals with BBU instantiation and the radio resource controller handles the radio resource allocation and beamforming.
- A dynamic radio and baseband resource allocation in the proposed architecture to improve radio resource utilization and IIoT users' satisfaction, by jointly optimizing the radio and baseband resource allocation, and the transmit beamforming, simultaneously.
  - A joint optimization of the PRB allocation, the RU-UE assignment, and the downlink transmit beamforming to minimize the NPC and maximize the ASR by employing the multi-objective optimization approach, under the proposed architecture. The reason for joint optimization is to consider user satisfaction when optimizing the resource allocation, which can be accomplished by minimizing NPC and maximizing ASR, simultaneously. The FDD channel estimation approach is used for imperfect channel state information (CSI) estimation to optimize the downlink transmit beamforming. The joint optimization problem is formulated as a non-convex and mixed-integer nonlinear problem (MINLP), which is NP-hard to solve. To solve the MINLP, we first relax the binary variables to obtain a continuous problem. Then, the obtained continuous problem is effectively solved by the successive convex approximation (SCA) approach.
  - A real-time RU-vBBU assignment to provide the required baseband resources so to handle the fluctuating traffic loads, under the proposed architecture. By exploiting the set of the active RUs obtained by the RU-UE assignment optimization, the number of vBBUs that need to be allocated to the active RUs is minimized. To this end, the MKP formulation is employed assuming the RUs as the objects and the vBBUs as the knapsacks capacity. The formulated MKP is solved, by decomposing the dual problems into sub-problems that are solved via the dual descent (DD) method.
- An extensive performance analysis to compare the proposed solution with its counterparts, in terms of the number of utilized radio and baseband resources, thereby highlighting the performance gains achieved through the proposed joint optimization of the radio and baseband resource allocation, and the transmit beamforming, simultaneously, as the solution to improve radio resource utilization as well as IIoT users' satisfaction.

The rest of the paper is organized as follows. In Section II, we discuss the related work. In Section III, we provide the system model. The problem formulation is presented in Section IV followed by the optimization solutions. The performance analysis and conclusions are presented in Sections V and VI, respectively.

Notation: $(.)^H$ denotes the Hermitan operator, $CN(0,x)$ represents the Gaussian distribution with zero mean and unit variance, and $\mathbb{C}$ is the complex set.

II. Related Work

In the current literature, the radio resource allocation in the C-RAN is mostly focused on cloud computing techniques. In [13], the authors proposed a dynamic resource allocation for C-RAN architecture exploiting the artificial intelligence. A prediction model based on the long short term memory (LSTM) is utilized for predicting the remote radio head (RRH) transmission. Finally, exploiting the predicted RRHs throughput, a genetic algorithm is used for resource allocation, aiming to improve resource utilization. In [14], the authors proposed a resource allocation for a virtual wireless network (VWN) in multiple-input multiple-output (MIMO)-aided C-RAN. The impact of pilot contamination error and pilot duration as the resource allocation optimization variable on the performance of VWN is studied. A two-step iterative algorithm is utilized, first, to adjust RRH, BBU, and backhaul parameters, then, to allocate the power to UEs. In [2], the authors proposed a resource management for IoT RAN with multicloud. A joint allocation of user, resource block, BS, cloud is studied. A heuristic scheduling algorithm is presented to optimize the resource allocation. In [15], a joint optimization of fronthaul and beamforming transmission in C-RAN architecture is proposed. The perfect and imperfect CSI scenarios are studied. The digital and analog beamforming are jointly optimized, since the precoding matrices affect the quantization noise. A block coordinate descent method is considered for the perfect CSI scenario and an iterative algorithm is utilized to obtain the efficient solution under imperfect CSI. In [16], a robust radio resource allocation for C-RAN architecture is proposed. The multiple-input and single-output (MISO) transmission mode with uncertain CSI is studied. An optimization problem is formulated for resource allocation, aiming to maximize the ASR.

In [17], a user selection and power minimization for C-RAN architecture is proposed. The signal-to-interference-and-noise ratio (SINR) required by user, RRH and user power constraints, and fronthaul capacity are considered to optimize the C-RAN performance. To this end, firstly, an algorithm based on the minimum-mean-square-error (MMSE) is proposed for user selection. Then, a reweighted-L_1 norm algorithm is employed for the NPC minimization. In [18], a coalition game method is proposed for minimizing the radio remote units (RRHs) transmission power while satisfying the target SINR in the C-RAN architecture. The beamformer is developed for a specific coalition structure in which RRHs greedily minimize the transmit power in the coalition without taking into account the interference to users in the other coalitions. In [19], the authors proposed an intelligent resource allocation based on Deep Q-Learning (DQL) in C-RAN architecture. A two-step solution is presented, first, to maximize the communication quality with the minimum number of the base stations, then, to maximize the resource utilization. The proposed DQL-based algorithm allocates the network resources in real-time depending on the decision-making structure obtained by slicing the existing channel conditions. In [20], the authors proposed an edge computing optimization for RRH-BBU assignment in cloud radio access network. A constrained resource allocation is modeled by the integer linear programming (ILP) formulation. They proposed two modified heuristic matroid based and Bi-Matching algorithms with low complexity to solve the formulated problem for the RRH-BBU assignment, aiming to reduce the fronthaul latency and the resource consumption. However, there is no investigation of the transmission signal to measure the achieved data rate by UEs.

Different from the existing efforts and motivated by maximizing resource utilization as well as satisfying users' satisfaction, we propose a joint optimization of the radio re-

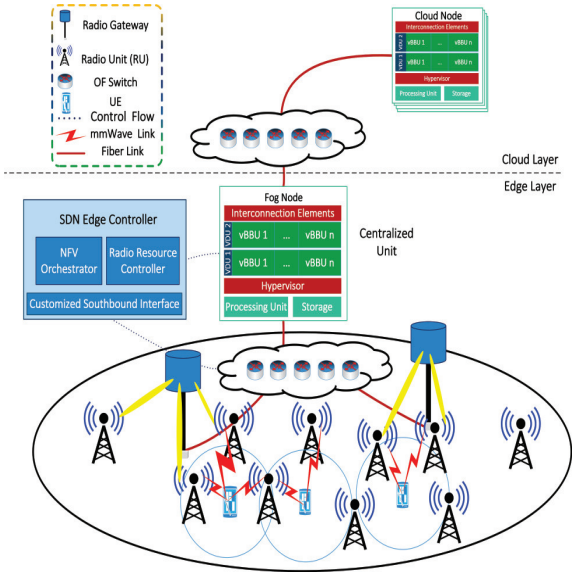


Fig. 1: An illustration of SDN based virtual Fog-RAN with 10 RUs.

source allocation and the transmit beamforming in SDN-based virtual Fog-RAN architecture for 5G-and-beyond wireless environments. To the best of our knowledge, this is the first work trying to optimize radio and baseband resource allocation, and the transmit beamforming, simultaneously, as the solution to improve radio resource utilization as well as users' satisfaction. Moreover, there is no solid effort in the literature to provide dynamic resource allocation in an SDN-based virtual Fog-RAN framework. The proposed SDN-based framework enables the Fog-RAN scalability by the NFV orchestrator and the radio resource controller, situated in the SDN edge controller, dealing with the baseband and radio resource allocation, respectively.

III. System Model

As shown in Fig. 1, we consider the downlink FDD transmission of an SDN-based virtual Fog-RAN containing I RUs denoted by the set $\boldsymbol{I} = \{1, ..., I\}$, where each RU is equipped with M antenna, and K UEs denoted by the set $\boldsymbol{K} = \{1, ..., K\}$. The Poisson process is utilized for scheduling the UEs' arrival with arrival rate λ and departure rate μ. We assume that UE k requests ζ_k PRB, which is the smallest unit of baseband resources that can be allocated to UEs for processing the baseband signals. Let us denote the set of $\boldsymbol{D} = \{1, ..., D\}$ the available PRBs, and $\gamma_{i,d}^{k} = 1$ means that the PRB d is allocated to the UE k on the RU i. Assuming user-centric clustering, a cluster of RUs, containing nearby RUs, is formed for each UE. To mitigate the computational complexity each UE can only be served by the RUs of its respective cluster. The clusters created may overlap since each RU can serve several UEs, concurrently. This overlap reduces the effect of interference on the UEs located at the cluster edge. Let $I_k \subseteq \boldsymbol{I}$ denotes the specific set of RUs serving the UE k, and the UEs covered by the RU i are represented by $K_i \subseteq \boldsymbol{K}$. Assuming coherent transmission of the signal to each UE by the RUs grouped in its respective cluster, the received baseband signal at UE k can be expressed as follows.

$$y_k = \sum_{i \in I_k} h_{i,k}^H w_{i,k} s_k + \sum_{l \in \boldsymbol{K}, l \neq k} \sum_{i \in I_l} h_{i,k}^H w_{i,l} s_l + z_k \qquad (1)$$

where $h_{i,k} \in \mathbb{C}^{M*1}$ and $w_{i,k} \in \mathbb{C}^{M*1}$ are the channel vector and the beamforming vector intended for UE k from RU i, respectively, s_k represents the signal with unit power intended for UE k, and $z_k \sim CN(0, \sigma^2)$ denotes the noise with power of σ^2.

A. Imperfect Channel Estimation

To obtain the imperfect CSI for UE k, we only estimate the CSI from $\forall i, i \in I_k$ because collecting the whole network's CSI is impractical in an ultra-dense RAN, while the large-scale fading coefficient is considered from $\forall i, i \in \boldsymbol{I}, i \notin I_k$. To this end, we exploit τ time slots from the available T time slots for estimating the channel state, while the remaining $T - \tau$ time slots are considered for data transmissions. Let $Q = [q_1, q_2, ..., q_\tau] \in \mathbb{C}^{\tau \times \tau}$ be the matrix of available pilot sequences with orthogonal column vectors. According to the FDD channel estimation approach, each RU i transmits the pilot sequences to all UEs $\forall k, k \in K_i$. Then, each UE estimates its channel based on the received signal and fed back to the RU for setting up the beamforming vector intended for the UE. Let us define the received pilot signal at UE k denoted by Y_k, which is calculated as follows.

$$Y_k = \sum_{i \in I_k} \sqrt{p_t} h_{i,k} Q_i^H + \sum_{l \in \boldsymbol{I}, l \notin I_k} \sqrt{p_t} h_{l,k} Q_i^H + N_k \qquad (2)$$

where $Q_i^H = \{q_{\pi_i}^1, ..., q_{\pi_i}^M\}$ denotes the transmitted pilot sequences by RU i, p_t represents the pilot transmit power, and N_k is the Gaussian noise vector. To facilitate channel estimation in ultra-dense RAN, the pilots should be reused among RUs as the number of RUs is much higher than the available pilot sequences. For reusing the pilots effectively, it is assumed the orthogonal pilot sequences must be allocated to the RUs within the same cluster, which is described by the constraint as follows.

$$q_{\pi_i} \neq q_{\pi_{i'}}, \forall i, i' \in \boldsymbol{I}, i \neq i', K_i \cap K_{i'} \neq 0 \qquad (3)$$

To satisfy constraint (3), we construct an undirected graph A, where any two RUs $i, i' \in \boldsymbol{I}$ that $i \neq i', K_i \cap K_{i'} \neq 0$ are linked together. Moreover, we limit the number of times each pilot can be reused to n_{max} to prevent the high training overhead. By taking into account the maximum number of reuse for each color n_{max}, the pilot allocation problem can be effectively solved as the coloring of the constructed graph A by Dsatur algorithm [21], since it offers low complexity and memory consumption. Let the set of RUs that reuse the same pilot as RU i does be denoted by $\Im_{\pi_i}$. By using the MMSE channel estimation [22], the estimate of channel $h_{i,k}$ can be expressed as follows.

$$\widehat{h}_{i,k} = \frac{\sqrt{p_t} \alpha_{i,k}}{\sum_{l \in \Im_{\pi_i}} p_t \alpha_{i,l+N_0}} Y_k Q_i^H \qquad (4)$$

where $\alpha_{i,k}, \forall i \in \boldsymbol{I}, i \notin \boldsymbol{I}_k$ represents the large-scale fading coefficient for RUs outside the cluster of UE k, and $N_0 = \sigma^2 / p_t$. According to the channel estimate vector $\widehat{h}_{i,k}$, the true channel vector $h_{i,k}$ is expressed as $h_{i,k} = \widehat{h}_{i,k} + \widetilde{h}_{i,k}$, where the error vector $\widetilde{h}_{i,k}$ demonstrates the CSI instability. Now, considering the imperfect channel estimation, the received baseband signal expressed in (1) can be rewritten as follows.

$$y_k = (\widehat{g}_{k,k}^H + \widetilde{g}_{k,k}^H) w_k s_k + \sum_{l \neq k, l \in \boldsymbol{K}} g_{l,k}^H w_l s_l + z_k \qquad (5)$$

where $w_k = \left[w_{i,k}^H, \forall i \in I_k\right]^H \in \mathbb{C}^{|I_k| M*1}$ denotes the transmit beamforming vectors from all RUs in I_k, $g_{l,k} = \left[h_{i,k}^H, \forall i \in I_l\right] \in \mathbb{C}^{|I_l| M*1}$ represents the aggregated channel

vector of all RUs in I_l to UE k, $\widehat{g}_{k,k}^H = \left[\widehat{h}_{i,k}^H, \forall i \in I_k\right]^H \in \mathbb{C}^{|I_k|M\times 1}$ represents the aggregated channel estimation vectors from all RUs in I_k to UE k, and $\widetilde{g}_{k,k}^H = \left[\widetilde{h}_{i,k}^H, \forall i \in I_k\right]^H \in \mathbb{C}^{|I_k|M\times 1}$ is the the aggregated error vector from all RUs in I_k to UE k. According to [14], the achievable data rate R_k considering the imperfect CSI can be expressed as follows.

$$R_k = \frac{T-\tau}{T} log_2\left(1+\Gamma_k^{Imperfect}\right) \tag{6}$$

$$\Gamma_k^{Imperfect} = \frac{\left|\widehat{g}_{k,k}^H w_k\right|^2}{\left|\widetilde{g}_{k,k}^H w_k\right|^2 + \sum_{l\neq k, l\in \boldsymbol{K}}\left|g_{l,k}^H w_l\right|^2 + \sigma^2} \tag{7}$$

where T denotes the total number of time slots in a coherence interval. Now, we define the ASR as the sum of the achievable data rate for all K UEs, which can be expressed as follows.

$$R^{total}(w) = \sum_{k\in\boldsymbol{K}} R_k(w_k) \tag{8}$$

where $\boldsymbol{w} = [w_1^H, w_2^H, ..., w_k^H]^H \in \mathbb{C}^{\boldsymbol{K}M\times 1}$.

B. Power Consumption

According to [23], the relation between the transmit power and the power consumption of the RU is nearly linear. Consequently, a linear approximation of the RU transmission power is justified to measure the power consumption of the RU. Therefore, the power consumption of RU i can be expressed as follows.

$$E_i^{ru} = \begin{cases} p_i^t(w) + \sum_{k\in I_k} p_t + E_i^{active}, & \text{if } b_i = 1 \\ E_i^{sleep}, & \text{if } b_i = 0 \end{cases} \tag{9}$$

where $p_i^t(w) = \sum_{k\in\boldsymbol{K}} p_{i,k}^t(w) = \sum_{k\in\boldsymbol{K}} ||w_{i,k}||^2$ represents the total transmit power at RU i, $b_i \in \{0,1\}, i \in \boldsymbol{I}$ is a binary variable to indicate the operational status of RU for which $b_i = 0$ implies the RU i is in the sleep state and $b_i = 1$ implies the RU i is in the active state, and E_i^{active} and E_i^{sleep} represent the amount of consuming power of RU i in active and sleep states, respectively. We denote by $E_{i,k}^f$ the power consumption for transferring the digital data from RU i to UE k. Let us define the association status between RU i and UE k by $v_{i,k} \in \{0,1\}, \forall i \in I_k$ and $k \in \boldsymbol{k}$, which $v_{i,k} = 1$ states that the UE k is served by the RU i and $v_{i,k} = 0$, otherwise. Therefore, the NPC can be expressed as follows.

$$E^{total}(w) = \eta\sum_{i\in\boldsymbol{I}}\sum_{k\in\boldsymbol{k}}||w_{i,k}||^2 + \sum_{i\in\boldsymbol{I}}\sum_{i\in I_k} b_i p_t + \sum_{i\in\boldsymbol{I}} b_i E_i^{active} + \sum_{i\in\boldsymbol{I}}(1-b_i)E_i^{sleep} + \sum_{i\in\boldsymbol{I}}\sum_{k\in\boldsymbol{k}} v_{i,k}E_{i,k}^f \tag{10}$$

where $\eta > 1$ is a constant for the efficiency of the power amplifier of RU.

IV. Problem Formulation and Solution

With the above analytic system modeling, we jointly optimize the radio resource allocation, and the downlink transmit beamforming. To this end, we split the joint optimization problem into two sub-problems and then solve both of them independently. The problem P_1 is a multi-objective optimization problem to optimize PRB allocation, UE-RU assignment, and transmit beamforming, aiming to minimize NPC and maximize ASR, simultaneously. The problem P_2 is a single objective optimization of RU-vBBU assignment.

A. Joint NPC and ASR Optimization Problem P_1

Motivated by [24], the optimization problem $\boldsymbol{P_1}$ can be written as follows.

$$\boldsymbol{P_1}: \min_{\mathbf{w},\mathbf{b},\mathbf{v},\gamma} \quad \Phi\frac{E^{total}(w)}{E^{Re.}} - (1-\Phi)\frac{R^{total}(w)}{R^{Re.}} \tag{11}$$

$$\text{s.t.}: \sum_{i\in\boldsymbol{I}}\sum_{d\in\boldsymbol{D}} v_{i,k}\gamma_{i,d}^k \leqslant \zeta_k, \forall k \in \boldsymbol{K} \tag{11a}$$

$$\forall (k,k') \in \boldsymbol{K}, \gamma_{i,d}^k + \gamma_{i,d}^{k'} \leqslant 1, \forall i \in \boldsymbol{I}, \forall d \in \boldsymbol{D} \tag{11b}$$

$$\Gamma_k(w) \geqslant \Gamma_k^{min}, \forall k \in \boldsymbol{K}, \tag{11c}$$

$$\sum_{k\in\boldsymbol{K}} p_{i,k}^t(w) \leqslant b_i p_i^{max}, \forall i \in \boldsymbol{I}, \tag{11d}$$

$$v_{i,k} \leqslant b_i, \forall i \in \boldsymbol{I}, \forall k \in \boldsymbol{K}, \tag{11e}$$

$$\sum_{i\in\boldsymbol{I}} v_{i,k} \geq 1, \forall k \in \boldsymbol{K}, \tag{11f}$$

$$||w_{i,k}||^2 \leqslant v_{i,k} p_{i,k}^t, \forall i \in \boldsymbol{I}, \forall k \in \boldsymbol{K}, \tag{11g}$$

$$p_{i,k}^t \leqslant v_{i,k} p_i^{max}, \forall i \in \boldsymbol{I}, \forall k \in \boldsymbol{K}, \tag{11h}$$

$$\sum_{k\in\boldsymbol{K}} v_{i,k} R_k(w_k) \leqslant \frac{Ca._i}{\xi_i}, \forall i \in \boldsymbol{I} \tag{11i}$$

$$b_i \in \{0,1\}, v_{i,k} \in \{0,1\}, \gamma_{i,d}^k \in \{0,1\}, \xi_i > 1. \tag{11j}$$

where $\Phi \in [0,1]$ determines the objective's weight. Note that if $\Phi = 1$, we obtain the minimization problem of NPC, and if $\Phi = 0$, we obtain the maximization problem of ASR. Due to the different magnitudes of the objective, $E^{total}(w)$ is divided by the reference value $E^{Re.}$, and $R^{total}(w)$ is divided by the reference value $R^{Re.}$ to ensure a consistent comparison. Constraint (11a) points out the number of PRBs allocated to each UE can not exceed the number demanded by the UE. Constraint (11b) stresses each PRB can only be allocated to one UE. Constraint (11c) emphasizes on the quality of service (QoS) requirements of UE k, where Γ_k^{min} denotes the minimum SINR needed for UE k. Constraint (11d) limits the total transmit power at RU i to the considered maximum power p_i^{max}. Constraints (11d) and (11e) ensure no power is emitted by the RU i, if $b_i = 0$. Constraint (11f) ensures at least one RU serves the UE k. Constraint (11g) emphasizes the transmit power to UE i by RU k should be zero, if $v_{i,k} = 0$. Constraint (11h) ensures the transmit power from the RU i to the UE k does not exceed p_i^{max}. Constraint (11i) enforces the fronthaul capacity required for a feasible transmission, where the capacity of the fronthaul link i needs to be ξ_i times greater than or equal to the achievable data rate at the RU i, for a defined fronthaul capacity factor ξ.

1) Solving Problem P_1 : Problem P_1 expressed in (11) is a MINLP due to the binary variables b, v, and γ, which is NP-hard to solve. With continuous relaxation of the variables b, v, and γ, the problem P_1 still remains non-convex, because the objective function in (11) and constraint (11i) are non-convex. Therefore, the problem P_1 is categorized as an MINLP non-convex problem that it is difficult to find a globally optimal solution for this problem. Next, we provide the procedures for optimally solving the intractable problem P_1.

First, we transform the binary variables by a continuous constraint to call continuous optimization. According to the well-known relaxation of binary variables given in [25], we can relax the binary variables as follows.

$$v_{i,k} \in \{0,1\}, \forall i,k \Leftrightarrow \sum_{i\in\boldsymbol{I},k\in\boldsymbol{K}} (v_{i,k}^2 - v_{i,k}) \geqslant 0, v_{i,k} \in [0,1] \tag{12}$$

$$b_i \in \{0,1\}, \forall i \Leftrightarrow \sum_{i \in \boldsymbol{I}} (b_i^2 - b_i) \geqslant 0, b_i \in [0,1] \quad (13)$$

$$\gamma_{i,d}^k \in \{0,1\}, \forall i,d \Leftrightarrow \sum_{i \in \boldsymbol{I}, d \in \boldsymbol{D}} ({\gamma_{i,d}^k}^2 - \gamma_{i,d}^k) \geqslant 0, \gamma_{i,d}^k \in [0,1] \quad (14)$$

The relaxation expressed in (12), (13), and (14) are justified by the fact that $v_{i,k}^2 - v_{i,k} < 0$ for $v_{i,k} \in [0,1]$, $b_i^2 - b_i < 0$ for $b_i \in [0,1]$, and ${\gamma_{i,d}^k}^2 - \gamma_{i,d}^k < 0$ for $\gamma_{i,d}^k \in [0,1]$, respectively. Thus, the problem P_1 can be rewritten as follows.

$$\min_{\Omega \in S_{Co.} \cap S_{NCo.}} \Phi \frac{E^{total}(w)}{E^{Re.}} - (1-\Phi)\frac{R^{total}(w)}{R^{Re.}} \quad \text{s.t. : (12), (13), (14).} \quad (15)$$

where $\Omega = \{w, b, v, \gamma\}$,and $S_{NCo.} = \{\Omega | (11i)\}$ and $S_{Co.} = \{\Omega | (11a) - (11h)\}$ are the set of convex and non-convex constraints of (15), respectively. Hence, $v_{i,k}$'s, b_i's, and $\gamma_{i,d}^k$'s are represented as continuous variables. Thus, the problem in (15) is a continuous non-convex one that can be solved by the SCA method [26]. The SCA utilizes an iterative algorithm for finding a solution of a non-convex problem by computing the convex approximation of the non-convex term. To determine an initial point of the iterative process, we apply the penalty method [27] that results in the regularization of the problem as follows.

$$\begin{aligned}\min_{\Omega \in S_{Co.} \cap S_{NCo.}} & \Phi \frac{E^{total}(w)}{E^{Re.}} - (1-\Phi)\frac{R^{total}(w)}{R^{Re.}} \\ & + \varphi_1 \sum_{i \in \boldsymbol{I}, k \in \boldsymbol{K}} (v_{i,k}^2 - v_{i,k}) + \varphi_2 \sum_{i \in \boldsymbol{I}} (b_i^2 - b_i) \\ & + \varphi_3 \sum_{i \in \boldsymbol{I}, k \in \boldsymbol{K}, d \in \boldsymbol{D}} ({\gamma_{i,d}^k}^2 - \gamma_{i,d}^k)\end{aligned} \quad (16)$$

where $\varphi_1 > 0$, $\varphi_2 > 0$, and $\varphi_3 > 0$ are the penalty parameters.

According to [26], the appropriate convex approximations of the objective function and the constraint (11i) are obtained as follows. Let us denote the non-convex objective function of problem P_1 by $F = R^{total}(w) = \sum_{k \in \boldsymbol{K}} R_k(w_k)$, and by $\tilde{F}(w_k, w_k(\nu))$ the convex approximation of the non-convex objective function F in the current iteration ν of the SCA algorithm around the feasible solution $w_k(\nu)$. The convex approximation of F can be obtained as follows.

$$\tilde{F}(w_k, w_k(\nu)) = \sum_{k \in \boldsymbol{K}} \tilde{F}_k(w_k, w_k(\nu)) + \bar{F}(w_k, w_k(\nu)) \quad (17)$$

where
$$\tilde{F}_k(w_k, w_k(\nu)) = R_k(w_k(\nu)) \quad (18)$$

In addition,
$$\bar{F}(w_k, w_k(\nu)) = \frac{\psi_{w_k}}{2} \left\| w_k - w_k(\nu) \right\|^2 \quad (19)$$

where ψ_{w_k} is an arbitrary positive constant. In (17), $\tilde{F}_k(w_k, w_k(\nu))$ is considered for the convexity of objective function and $\bar{F}(w_k, w_k(\nu))$ provides the strong convexity.

To compute the convex upper bound for the constraint (11i), we first rewrite the constraint (11i) as follows.

$$g = \sum_{k \in \boldsymbol{K}} v_{i,k} R_k(w_k) - \frac{Ca._i}{\xi_i} \leqslant 0, \forall i \in \boldsymbol{I} \quad (20)$$

The constraint g is a non-convex constraint with a difference convex (DC) structure. Let $\tilde{g}(w_k, w_k(\nu))$ be the convex upper approximation of constraint g in the current iteration ν of the SCA algorithm around the feasible solution $w_k(\nu)$. By making the concave part of g linear and leaving the convex part of g unchanged, we can obtain $\tilde{g}(w_k, w_k(\nu))$ as follows.

$$\begin{aligned}\tilde{g}(w_k, w_k(\nu)) = & -\frac{Ca._i}{\xi_i} + \sum_{k \in \boldsymbol{K}} v_{i,k} R_k(w_k) \\ & - v_{i,k} \nabla_{w_k} R_k(w_k(\nu))(w_k - w_k(\nu)) \geqslant g\end{aligned} \quad (21)$$

where $\nabla_{w_k} R_k(w_k(\nu))$ denotes the gradient of R_k with respect to the feasible solution $w_k(\nu)$.

Having the convex approximation of the objective function (11) and the constraint (11i), instead of solving P_1, we solve the given problem by the SCA algorithm, as follows.

$$\begin{aligned}\min_{\Omega \in S_{Co.}} & \Phi \frac{E^{total}(w)}{E^{Re.}} - (1-\Phi)\frac{\tilde{F}(w_k, w_k(\nu))}{R^{Re.}} \\ & + \varphi_1 \sum_{i \in \boldsymbol{I}, k \in \boldsymbol{K}} (v_{i,k}^2 - v_{i,k}) + \varphi_2 \sum_{i \in \boldsymbol{I}} (b_i^2 - b_i) \\ & + \varphi_3 \sum_{i \in \boldsymbol{I}, k \in \boldsymbol{K}, d \in \boldsymbol{D}} ({\gamma_{i,d}^k}^2 - \gamma_{i,d}^k) \\ & \text{s.t. :(20).}\end{aligned} \quad (22)$$

The SCA based algorithm for solving the problem P_1 is described in algorithm 1. The idea is to use convex problems iteratively to approximate the original non-convex problem. In this algorithm, ν denotes the iteration parameter and $w(0), b(0), v(0), \gamma(0)$ denote the initial points chosen from the feasible region. We solve the convex approximate problem of (22) in each iteration of algorithm 1. The iterative procedure is repeated until the stopping criteria is satisfied, where $|w^*(\nu+1) - w^*(\nu)| < \delta_1$, $|b^*(\nu+1) - b^*(\nu)| < \delta_2$, $|v^*(\nu+1) - v^*(\nu)| < \delta_3$, and $|\gamma^*(\nu+1) - \gamma^*(\nu)| < \delta_4$. The values of δ_1, δ_2, δ_3 and δ_4 are the differences between two successive iterations of the respective objective function values.

Algorithm 1: The SCA-based algorithm

$\nu \leftarrow 0$; Choose initial values:w$(\nu), b(\nu), v(\nu), \gamma(\nu)$;
repeat
  Compute problem in (22) with $w(\nu), b(\nu), v(\nu), \gamma(\nu)$ to obtain w^*, b^*, v^*, γ^*;
  Update $w(\nu+1) \leftarrow w^*, b(\nu+1) \leftarrow b^*$, $v(\nu+1) \leftarrow v^* \gamma(\nu+1) \leftarrow \gamma^*$;
  $\nu \leftarrow \nu + 1$;
until *stopping criteria is satisfied*;
Return $\Omega^* = \{w^*, b^*, v^*, \gamma^*\}$

B. RU-vBBU Assignment Problem P_2

In the conventional C-RAN architecture, one vBBU is assigned to one particular RU for handling its traffic loads. This RU-vBBU assignment can not efficiently utilize the resources, because the traffic loads of the RUs are not equal. The problem P_2 aims to minimize the number of vBBUs assigned to the active RUs obtained in the problem P_1. To achieve this objective, we employ the MKP [29], where the RUs are considered as the objects and the vBBUs as the knapsacks. In each epoch of time, an RU-vBBU assignment is dynamically performed considering the traffic load fluctuation. Let N denotes the optimal number of vBBUs required to handle the set of active RUs, which can be calculated as follows.

$$N = \left\lceil \frac{\sum_{i \in \boldsymbol{I}} \sum_{k \in \boldsymbol{K}} \sum_{d \in \boldsymbol{D}} b_i v_{i,k} \gamma_{i,d}^k}{D} \right\rceil \quad (23)$$

where D represents the total number of PRBs. We introduce a binary variable $r_{j,i} \in \{0,1\}$ to indicate whether the RU i is assigned to the vBBU j or not. We define the profit of assigning the RU i and the vBBU j as the satisfaction ratio of the RU i denoted by $\rho_{j,i}$, which is calculated by

$$\rho_{j,i} = \frac{c_{j,i}}{\sum_{k \in K_i} \zeta_k} \tag{24}$$

where $c_{j,i}$ denotes the number of the PRBs resources offered by the vBBU j and consumed by the RU i that is calculated by

$$c_{j,i} = \sum_{k \in K_i} \sum_{d \in \boldsymbol{D}} r_{j,i} b_i v_{i,k} \gamma_{i,d}^k \tag{25}$$

Assuming each vBBU can entirely handle a fully loaded RU, we formulate the MKP for RU-vBBU assignment as follows.

$$\boldsymbol{P_2}: \max_r \quad \sum_{j=1}^{N} \sum_{i=1}^{I} \rho_{j,i} r_{j,i} \tag{26}$$

$$\text{s.t.}: \quad \sum_{j=1}^{N} \sum_{i=1}^{I} c_{j,i} r_{j,i} \leqslant D, \tag{26a}$$

$$\sum_{i \in Z} r_{j,i} \leqslant 1, \forall j \in \{1,...,N\}, Z \subseteq \{1,...,I\} \tag{26b}$$

$$r_{j,i} \in \{0,1\}, i \in \{1,...,I\}, j \in \{1,...,N\}. \tag{26c}$$

Constraint (26a) ensures the vBBU j can provide the required resources of the assigned RUs, and constraint (26b) limits the vBBU resource consumption per RU.

1) Solving Problem P_2 : The MKP formulated in (26) is an integer linear problem (ILP), which can be solved by decomposing the dual problems into sub-problems and exploiting the DD method to solve them. Let us define by $\lambda_j, j \in \{1,...,N\}$ the Lagrangian multipliers for the global constraints. Then, the problem P_2 in (26) can be written as

$$\max_r \quad \sum_{j=1}^{N} \sum_{i=1}^{I} \rho_{j,i} r_{j,i} - \sum_{j=1}^{N} \lambda_j \Big(\sum_{j=1}^{N} \sum_{i=1}^{I} c_{j,i} r_{j,i} - D \Big) \tag{27}$$

$$\text{s.t.:} \sum_{i \in Z} r_{j,i} \leqslant 1, \forall j \in \{1,...,N\}, Z \subseteq \{1,...,I\} \tag{27a}$$

$$r_{j,i} \in \{0,1\}, i \in \{1,...,I\}, j \in \{1,...,N\}. \tag{27b}$$

satisfying the optimality conditions given below.

$$\lambda_j \Big(\sum_{j=1}^{N} \sum_{i=1}^{I} c_{j,i} r_{j,i} - D \Big) = 0, \lambda_j \geqslant 0 \tag{28}$$

$$\sum_{j=1}^{N} \sum_{i=1}^{I} c_{j,i} r_{j,i} - D \leqslant 0 \tag{29}$$

The problem in (27) is then decomposed into a set of sub-problems in each epoch of time, one for each vBBU $j, j \in \{1,...,N\}$, as follows.

$$\max_r \quad \sum_{i=1}^{I} \rho_{j,i} r_{j,i} - \lambda_j \sum_{i=1}^{I} c_{j,i} r_{j,i} \tag{30}$$

$$\text{s.t.}: \quad \sum_{i \in Z} r_{j,i} \leqslant 1, Z \subseteq \{1,...,I\} \tag{30a}$$

$$r_{j,i} \in \{0,1\}, i \in \{1,...,I\}. \tag{30b}$$

The dual decomposition in (30) offers a distributed approach to solve the formulated MKP by alternating between concurrently solving the independent sub-problems for each given λ_j and updating $_j$, while getting converged to the solution r. Let us construct a directed acyclic graph (DAG) G for $\{Z | Z \subseteq \boldsymbol{I}\}$, where a directed edge is formed between (z, z') for all $z, z' \in Z$ if $K_z \cap K_{z'} \neq 0$ and $\widetilde{\rho}_{j,z} < \widetilde{\rho}_{j,z'}$. The adjusted profit of the RU $z \in Z$, assigned to vBBU j, is calculated as follows.

$$\widetilde{\rho}_{j,z} = \rho_{j,z} - \frac{c_{j,i}}{D} \tag{31}$$

Algorithm 2: Distributed Dual Descent Method

input : $\rho_{j,i}, c_{j,i}, G$

Function $\boldsymbol{Map}\,(\rho_{j,i}, c_{j,i})$ **:**
- **for** ***each*** *($Z \subseteq \boldsymbol{I}$ in topological order of G)* **do**
  - $l \leftarrow |Z|$; $\forall i : r_{j,i} = 0$; $c_j = 0$;
  - **repeat**
    - **if** $(c_j + c_{j,l} \leq D)$ **then**
      - $r_{j,l} \leftarrow 1$;
      - $c_j \leftarrow c_j + c_{j,l}$;
    - $l \leftarrow l\text{-}1$;
  - **until** $(c_j \geq D)$;
- **for** ***each*** *(j in the set $\{1,...,N\}$)* **do**
  - $\varsigma_{j,i} \leftarrow \sum_{i=1}^{I} c_{j,i} r_{j,i}$
  - ***Return***$(j, \varsigma_{j,i})$

Function $\boldsymbol{Reduce}\,(j, \varsigma_{j,i})$ **:**
- **return** $\sum_{j=1}^{N} \varsigma_{j,i}$

Function $\boldsymbol{Main}$**:**
- ***Initialize*** λ^0;
- **for** $t \leftarrow 1$ **to** $t-1$ **do**
  - **for** ***each*** *($j \in \{1,...,N\}$)* ***in parallel*** **do**
    - *Map*$(\rho_{j,i}, c_{j,i})$
  - **for** ***each*** *($j \in \{1,...,N\}$)* ***in parallel*** **do**
    - $W_j \leftarrow$ *Reduce*$(j, \varsigma_{j,i})$
  - **for** ***each*** *($j \in \{1,...,N\}$)* **do**
    - $\lambda_j^{t+1} \leftarrow \max\Big(\lambda_j^t + \psi(W_j - D), 0\Big)$
  - **if** *(λ_t has converged)* **then**
    - **return** λ^t
- **return** $\lambda^{\{0,..,t-1,t\}}$

The MKP can then be solved by the distributed DD method, introduced in algorithm 2, according to the MapReduce model [30]. In each epoch of times, the solutions $r_{j,i}$ for the sub-problems are firstly computed independently for each vBBU, by traversing the constructed graph G in a topological order via mappers. The RUs are sorted in a non-decreasing order based on the adjusted profit. Thus, none RU has less adjusted profit than its preceding one. In this way, the RUs that serve the same UE can be assigned to the same vBBU. Starting from the lowest level of the DAG, the algorithm selects the RUs with the highest adjusted profit until the sum of the consuming PRBs reaches D. The $r_{j,i}$ for the chosen RUs is assigned with the value of 1, and for the rest of the RUs members of Z is assigned with the value of 0. Then, each mapper returns N values $\{\varsigma_{j,i} = \sum_{i=1}^{I} c_{j,i} r_{j,i} | j \in \{1,...,N\}\}$ corresponding to each vBBU. Next, the reducers sum all the values returned for a particular vBBU, $W_j = \sum_{j=1}^{N} \varsigma_{j,i}$. In the end, a master node updates the multipliers λ_j as follows.

$$\lambda_j^{t+1} = \max\Big(\lambda_j^t + \psi\Big(\sum_{j=1}^{N} \sum_{i=1}^{I} c_{j,i} r_{j,i} - D\Big), 0\Big) \tag{32}$$

where ψ is the step size.

Proposition 1: Algorithm 2 optimally solves the sub-problem in (30).

Proof: Let $r_{j,i}$ be an alternative of $r^*_{j,i}$ that satisfies the constraints (30a) and (30b). Then, the first node may also be identified with distinct chosen RUs in the topological order of the DAG. Therefore, there is a pair of RUs i and i', where the adjusted profit of i is not less than i', while $r^*_{j,i} = 1$, $r^*_{j,i'} = 0$ and $r_{j,i} = 0$, $r_{j,i'} = 1$. The constraints (30a) and (30b) are still satisfied when we set $r_{j,i} = 1$, $r_{j,i'} = 0$, since the objective value of (30) remains unchanged and both RUs i and i' exist later in the topological order.

C. Running Time Analysis

The overall running time of the proposed solution mainly depends on solving the problems P_1 and P_2. According to the $2^{I+IK+ID}$ combination of binary variables $b_i, v_{i,k}$, and $\gamma^k_{i,d}$ and $3IK + 2I + 3K + ID$ constraints in (11a):(11j), the worst-case running time of our algorithm for P_1 is evaluated as $O\big(2^{IK+K}(K^2I^2D)\big)$. However, in order to obtain a practical running time, with the expected massive number of UEs in IIoT wireless environments, several low-complexity methods are proposed to find the feasible solution to P_1. To this end, the running time is remarkably reduced by applying a continuous relaxation on the binary variables. Moreover, an SCA method is employed to relax the non-convexity of problem P_1. Furthermore, the RU i is not selected if it does not serve any UE (i.e., K_i is zero) to involve the sparsity method for the RUs selection. Thus, the worst-case running time of our algorithm for P_1 is reduced to $O\big(K^2I^2D\big)$, by using continuous relaxation, convexification, and sparsity methods, resulting in the proposed solution converging rapidly. Fig. 2 indicates the convergence behaviour of the proposed SCA solution for different number of UEs. It can be observed that the proposed solution provides fast convergence to the optimal solution. Also, this convergence time does not increase with the number of UEs. Moreover, the worst-case running time of our algorithm for P_2 is evaluated as $O(N^2I)$.

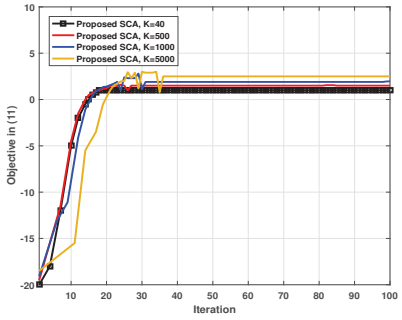


Fig. 2: Convergence behaviour of the proposed SCA solution for different number of UEs.

V. Performance Analysis

In this Section, we present the simulation results, using the Matlab tool, for evaluating the performance of the proposed joint optimization, where the simulation parameters are listed in Table I. Moreover, we assume the Rayleigh fading channel model with zero mean and unit variance and the path loss $148.1{+}37.6log_{10}dis.$, where $dis.$ is the distance (km) between a particular RU and UE. Furthermore, We use the Poisson model with a fixed arrival rate of $\lambda = 5$ and departure rate of $\mu = 0.1$ for UEs, where the demand of UEs follow the uniform distribution $u = (1, 10)$.

We study the performance of the proposed SCA solution under the three scenarios of (1) imperfect intra-CSI, (2) perfect intra-CSI, and (3) complete CSI. In particular, the imperfect intra-CSI scenario estimates the imperfect CSI from all the RUs within the UE's cluster and considers the long-scale fading coefficient for RUs outside the cluster, while the perfect intra-CSI scenario estimates the perfect CSI without the error estimation. Moreover, the complete CSI scenario estimates the entire network's CSI for each UE, so the orthogonal pilot is needed for each RU, since the pilot can not be reused.

TABLE I: SIMULATION PARAMETERS.

Parameter	Value	Parameter	Value
I	50	σ^2	-174 dBm/Hz
K	40	E^f	1 dBW
M	2	Ca.	500 b/s/Hz
D	100	ξ	10
p_i^{max}	10 dBW	$\delta_1, \delta_2, \delta_3, \delta_4$	10^{-5}
T	100	E^{active}	11 dBW
p_t	10 dBW	E^{sleep}	1 dBW
$\mathrm{E}^{Re.}$	1 b/s/Hz	$\mathrm{R}^{Re.}$	0 dBW

First, we investigate the impact of the cluster size L on the ASR and the NPC. The user-centric clustering is applied to limit the number of RUs serving each UE, which reduces the computational complexity. We consider $|I_k| = L, \forall k \in \boldsymbol{K}$ so to have a fixed cluster for each UE, because the large-scale fading coefficient shifts slowly. Fig. 3(a) illustrates the ASR versus the cluster size for the different CSI scenarios. It can be observed the ASR for the imperfect and perfect intra-CSI scenarios firstly increases by enhancing the cluster size until $L = 10$ with $n_{max} = 3$ and until $L = 12$ with $n_{max} = 6$. The larger cluster size allows more RUs to coherently transmit the baseband signal to each UE, so the UEs receive a stronger superposition of signals that increases the ASR. By enhancing the cluster size beyond 10 with $n_{max} = 3$ and beyond 12 with $n_{max} = 6$, the ASR declines. This is because enhancing the cluster size allows more RUs to be grouped in each cluster that increases the required orthogonal pilots. Enhancing the orthogonal pilots reduces the remaining time slots for data transmission, so the ASR declines. Therefore, the cluster size and the maximum pilot reuse need to be carefully considered to avoid complexity and high pilot overhead. To achieve adequate performance and low complexity, we consider the cluster size be $L = 8$ and the maximum pilot reuse be $n_{max} = 3$ for the rest of the simulations. The reason for this choice is that the improvement of the ASR by $n_{max} = 6$ is negligible. Moreover, the ASR with $n_{max} = 3$ after cluster size $L = 8$ grows marginally. It can be observed in Fig. 3(a), the ASR for the complete CSI scenario increases along with the cluster size, unlike the imperfect and perfect intra-CSI scenarios. This is because the number of required orthogonal pilots in the complete CSI scenario is equal to the number of RUs and is independent of the cluster size; thus, a larger n_{max} cannot affect the ASR. Also, as shown in Fig. 3(a), the imperfect intra-CSI scenario offers better ASR compared to the perfect intra-CSI scenario, because it measures the imperfect CSI by considering the error estimation. Fig. 4(a) shows the NPC versus the cluster size L with maximum pilot reuse $n_{max} = 3$. It can be observed the NPC increases along with the cluster size. The reason is that more RUs can transmit power to each UE with larger cluster size, so the NPC grows. It is observed the growth slope of NPC for the imperfect and perfect intra-CSI scenarios is slightly reduced beyond $L = 10$. The reason for the reduction in growth slope is that more orthogonal pilots are required when the cluster size is beyond 10. Thus, the number of remaining slots for data transmission is reduced, resulted in lower power consumption. However, transmitting more orthogonal pilots still slightly raises the power consumption. It can be observed the perfect intra-CSI scenario consumes lower power than the imperfect intra-CSI scenario. The reason is the imperfect intra-CSI scenario feds back imperfect CSI by considering the error estimation. Thus, RUs transmit more power to provide the

requested demands. The imperfect intra-CSI offers much lower power consumption than the complete CSI scenario, since the complete CSI scenario needs to obtain the entire network's CSI by transmitting the orthogonal pilots per RU.

Furthermore, we study the impact of the minimum SINR Γ^{min} on the ASR and the NPC. Fig. 3(b) presents the ASR versus the minimum SINR Γ^{min} for the imperfect intra-CSI, perfect intra-CSI, and complete CSI scenarios. The ASR increases along with the minimum SINR Γ^{min}. Also, the ASR increases as p^{max} is increased, because RUs can provide a stronger signal to UEs. Moreover, the impact of the error estimation on the ASR is also verified by comparing the curves of the imperfect and perfect intra-CSI scenarios when the minimum SINR is large. It is obvious that the RUs need to transmit more power to meet the higher minimum SINR required by the UEs, so this increases the interference. Thus, employing the error estimation in the imperfect intra-CSI scenario causes better channel estimation resulted in better ASR, and consequently larger NPC. Fig. 4(b) presents the NPC versus the minimum SINR Γ^{min} for the imperfect intra-CSI, perfect intra-CSI, and complete CSI scenarios. It can be observed the NPC increases as the minimum SINR Γ^{min} increases, because the RUs transmit more power to meet the minimum SINR. The NPC increases sharply for low Γ^{min}, while it slightly increases beyond $\Gamma^{min} = 4$. This is because fewer transmitters need to be activated to provide a minimum SINR of more than 4, since the active RUs can provide the required SINR. As expected, the NPC increases as the maximum RU transmit power p^{max} increases, because the RUs can transmit higher power. Fig. 3(c) demonstrates the ASR versus the number of UEs for the imperfect intra-CSI, perfect intra-CSI, and complete CSI scenarios. As expected, the ASR increases along with the number of UEs. It is observed the error estimation has a higher impact on ASR with a higher number of UEs and a larger p^{max}, because the interference grows by increasing the number of UEs and p^{max}, so the error estimation causes better channel estimation. Moreover, Fig. 4(c) illustrates the NPC versus the number of UEs for the imperfect intra-CSI, perfect intra-CSI, and complete SCI scenarios. As expected, the NPC increases along with the number of UEs, since more RUs need to be activated to serve the UEs. Moreover, it can be observed the NPC increases as the maximum RU transmit power p_{max} increases, because the RUs can transmit higher power.

Moreover, we investigate the trade-offs between ASR and NPC in the proposed SCA algorithm, as well as we provide the respective comparisons with the reweighted-L_1 norm algorithm introduced in [17] and the coalitional game algorithm introduced in [18]. The maximum ASR is obtained by setting $\Phi = 0$, while $\Phi = 1$ minimizes the NPC. Moreover, different trade-offs for ASR maximization and NPC minimization are obtained by altering Φ. Fig. 5(a) represents the trade-offs between ASR and NPC. As expected, the ASR increases along with the NPC. The higher power consumption means more power is transmitted to UEs leading to an increase of the ASR. It is observed the ASR improvement is negligible for a high NPC. This is because the ASR improvement depends on the number of serving RUs to each UE, which is limited by the cluster size. Thus, increasing the number of the active RUs that enhances the NPC does not necessarily improve the ASR significantly. Also, it can be observed that the proposed algorithm outperforms the reweighted-L_1 norm algorithm and coalitional game algorithm in terms of ASR and NPC.

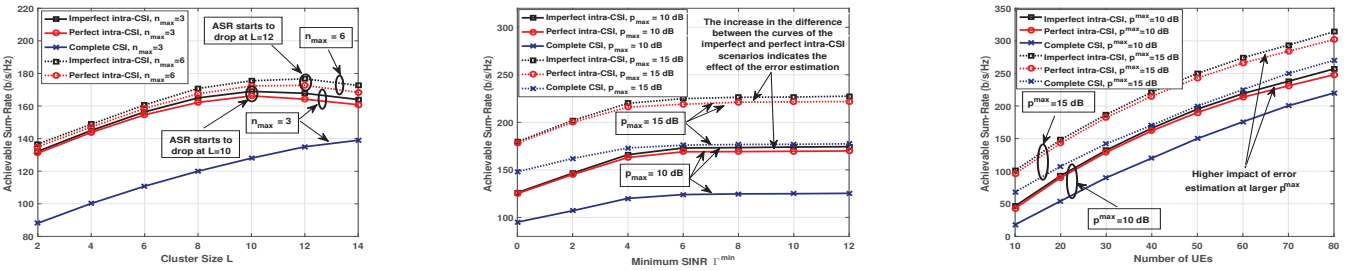
To gain further insights into the joint optimization, we study the impact of the arrival rate λ. Fig. 5(b) illustrates the UE's satisfaction ratio, defined as the ratio of the allocated PRBs to the requested PRBs, versus the UE's arrival rate λ for the proposed SCA solution with different maximum RU transmit power p^{max} and uniform distribution of UE's PRB demand. We further provide the respective comparisons with the reweighted-L_1 norm algorithm [17] and the coalitional game algorithm [18]. It can be observed the satisfaction ratio for the proposed SCA solution decreases as $\lambda = 5$ increases if more RUs are not activated. The reason is that the number of requested PRBs increases with increased λ, so the current number of active RUs can not provide all the requested PRBs. Moreover, the satisfaction ratio increases by increasing p^{max}, because the ASR increases. The proposed SCA clearly outperforms the reweighted-L_1 norm algorithm and the coalitional game algorithm in terms of satisfaction ratio, because of its low computational complexity, as clearly stated in Section IV.C, which can provide a high satisfaction ratio even at high arrival rates.

Fig. 5(c) demonstrates the number of active RUs and vBBUs versus the UE's arrival rate λ. It can be observed the number of active RUs increases along with λ as a larger number of PRB resources are needed when the number of UEs increases. The proposed SCA solution only switches on 40% of the RUs when the arrival rate $\lambda = 1$ and 74% of the RUs when the arrival rate $\lambda = 10$, causing a significant saving in OPEX, while in the conventional C-RAN the RUs are always active. Moreover, as it is clearly observed, the proposed SCA solution provides better radio resources savings, compared with the reweighted-L_1 norm and coalitional game solutions. The number of vBBU units is independent of the arrival rate λ for the conventional C-RAN. The number of vBBUs is equal to 50 at any arrival rate, since the RUs are always active. We compare our proposed MKP solution for RU-vBBU assignment with the Matroid based and Bi-Matching solutions introduced in [20]. It can be observed the proposed RU-vBBU assignment based on the MKP solution obtains significant savings in vBBUs compared to the one-to-one, Matroid based, and Bi-Matching assignments. In the maximum arrival rate $\lambda = 10$, only 30% of the vBBU resources are allocated by our proposed MKP solution, whereas, the one-to-one, Matroid based, and Bi-Matching solutions use 74%, 56%, and 38% of the vBBU resources, respectively. As a conclusion, from Fig.4(c), the proposed SCA solution with the MKP assignment, when $\lambda = 10$, provides 22% and 27% more saving in radio resources than the reweighted-L_1 norm and coalitional game solutions,respectively, and 46% and 21% more saving in baseband resources than the Matroid based and Bi-Matching solutions, respectively.

Fig. 6 illustrates the objective in (11) versus the objective's weight Φ. It is observed that the objective first decreases but then increases as the parameter Φ increases. According to (11), NPC increases when Φ increases that results in the objective reduction. After some point, the objective increases when the first term in (11) dominates the second term. Fig. 6 again shows that the proposed SCA outperforms the reweighted-L_1 norm and the coalitional game algorithms. Fig. 7 demonstrates the convergence behaviour for the proposed SCA, the reweighted-L_1 norm and the coalitional game algorithms. We observe the proposed SCA algorithm only needs few iterations to be converged, whereas, the reweighted-L_1 norm and coalitional game algorithms requires more iterations to converge. Moreover, the proposed SCA converges to a smaller objective compared with its counterparts.

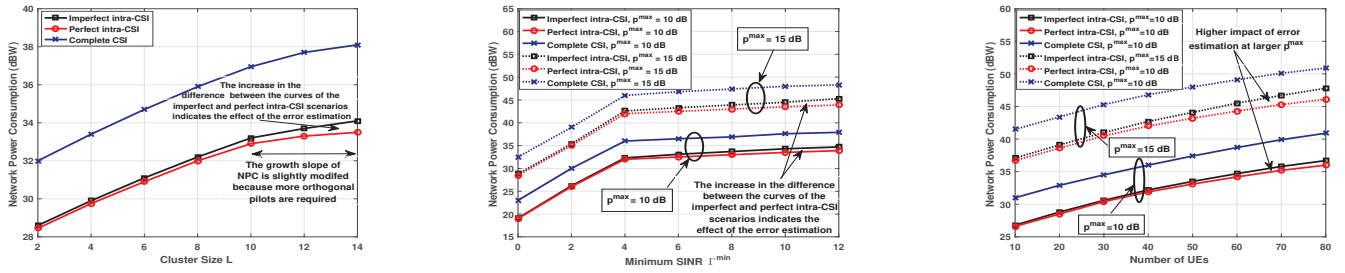
VI. Conclusions

A joint radio resource allocation and downlink transmit beamforming optimization in SDN-based virtual Fog-RAN 5G-and-beyond wireless environments is proposed to support the massive number of devices in future IIoT networks. The major objective is to improve the radio resource utilization and the IIoT users' satisfaction by maximizing the ASR and



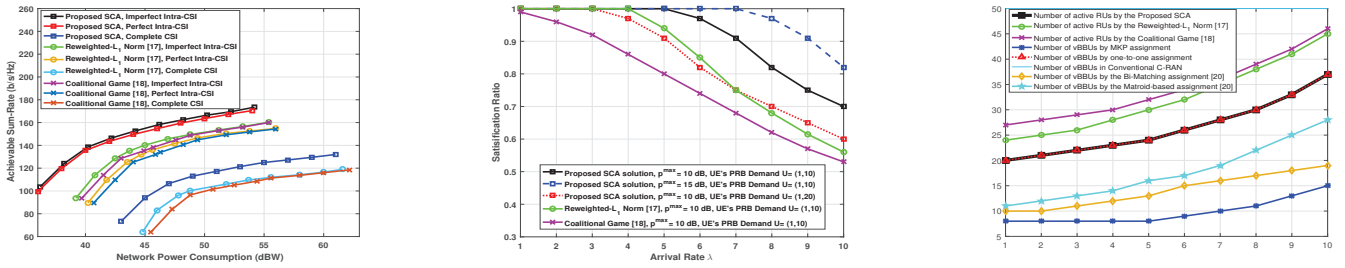
(a) (b) (c)

Fig. 3: (a) ASR vs. cluster size L, (b) ASR vs. minimum SINR Γ^{min}, (c) ASR vs. number of UEs, with M=2, I=50



(a) (b) (c)

Fig. 4: (a) NPC vs. cluster size L, (b) NPC vs. minimum SINR Γ^{min} , (c) NPC vs. number of UEs, with M=2, I=50.



(a) (b) (c)

Fig. 5: (a) Trade-offs between ASR and NPC, (b) Satisfaction ratio vs. λ, (c) Number of units vs. λ, with M=2, I=50

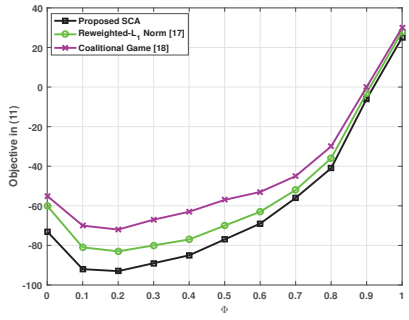


Fig. 6: Objective in (11) versus parameter Φ.

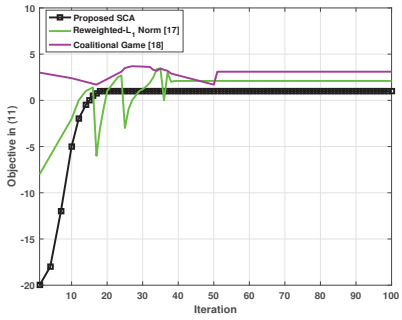


Fig. 7: Convergence behaviour for different algorithms.

minimizing the NPC. To this end, the PRB resource allocation, the UE-RU assignment, the transmit beamforming, and the RU-vBBU assignment are studied within a single framework. We first formulated a MINLP problem to jointly maximize the ASR and minimize the NPC that is solved by the SCA method. Then, exploiting the set of active RUs minimized in the previous optimization problem, we formulated the RU-vBBU assignment by the MKP formulation that is solved by decomposing the dual problems into sub-problems and employing the DD method to solve them. Through an extensive performance analysis, we demonstrated that our proposed SCA solution outperforms its counterparts in terms of the ASR and the NPC. Moreover, the proposed SCA solution based MKP assignment offers important savings in baseband and radio resources, outperforming its counterparts as well. Therefore, the performance gains achieved validate the usefulness of the proposed joint optimization as a promising solution for handling the huge traffic loads arisen from the massive number of devices in future IIoT networks. As a future direction, the applicability of the proposed solution for the user centric cell-

free Fog-RAN can be investigated. Moreover, the impact of the multiple central units with radio stripes technology can be studied.

REFERENCES

[1] S.K. Sharma, X. Wang, "Towards Massive Machine Type Communications in Ultra-Dense Cellular IoT Networks: Current Issues and Machine Learning-Assisted Solutions," IEEE Journal on Communication and Surveys and Tutorials, 2019.

[2] M. Awais, A. Ahmed, S. A. Ali, M. Naeem, W. Ejaz and A. Anpalagan, "Resource Management in Multicloud IoT Radio Access Network," in IEEE Internet of Things Journal, vol. 6, April 2019.

[3] D. Pliatsios, P. Sarigiannidis, S. Goudos, and G.K. Karagiannidis, "Realizing 5G vision through Cloud RAN: technologies, challenges, and trends," J. Wir. Com. Net., vol. 136, May 2018.

[4] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat and H. Dai, "A Survey on Low Latency Towards 5G: RAN, Core Network and Caching Solutions," in IEEE Communications Surveys Tutorials, vol. 20, no. 4, 2018.

[5] R. I. Tinini, D. M. Batista, G. B. Figueiredo, M. Tornatore and B. Mukherjee, "Low-latency and energy-efficient BBU placement and VPON formation in virtualized cloud-fog RAN," in IEEE/OSA Journal of Optical Communications and Networking,2019.

[6] M. Peng, S. Yan, K. Zhang and C. Wang, "Fog-computing-based radio access networks: issues and challenges," in IEEE Network, vol. 30, no. 4, pp. 46-53, July-August 2016.

[7] M. A. Habibi, M. Nasimi, B. Han and H. D. Schotten, "A Comprehensive Survey of RAN Architectures Toward 5G Mobile Communication System," in IEEE Access, vol. 7, pp. 70371-70421, 2019.

[8] R. Yu, G. Xue, M. Bennis, X. Chen, and Z. Han, "HSDRAN: Hierarchical software-defined radio access network for distributed optimization," IEEE Trans. Veh. Technol., vol. 67, no. 9, pp. 8623–8636, Sep. 2018.

[9] H. Hawilo, A. Shami, M. Mirahmadi, and R. Asal, "NFV: state of the art, challenges, and implementation in next generation mobile networks (vEPC)," IEEE Netw., vol. 28, no. 6, pp. 18–26, 2014.

[10] Tayyaba, S.K., Shah, M.A. Resource allocation in SDN based 5G cellular networks. Peer-to-Peer Netw. Appl. 12, 514–538 (2019).

[11] R.T. Rodoshi, T. Kim, W. Choi, "Resource Management in Cloud Radio Access Network: Conventional and New Approaches," J. Sensors, 2020.

[12] W. Ejaz, S.K. Sharma, S. Saadat, M. Naeem, A. Anpalagan, and N.A. Chughtai, "A comprehensive survey on resource allocation for CRAN in 5G and beyond networks," J. Net. Compu. App., vol. 160, 2020.

[13] W. Chien, C. Lai and H. Chao, "Dynamic Resource Prediction and Allocation in C-RAN With Edge Artificial Intelligence," in IEEE Transactions on Industrial Informatics, vol. 15, no. 7, July 2019.

[14] S. Parsaeefard, R. Dawadi, M. Derakhshani, T. Le-Ngoc and M. Baghani, "Dynamic Resource Allocation for Virtualized Wireless Networks in Massive-MIMO-Aided and Fronthaul-Limited C-RAN," in IEEE Transactions on Vehicular Technology, vol. 66, no. 10, Oct. 2017.

[15] J. Kim, S. Park, O. Simeone, I. Lee and S. Shamai Shitz, "Joint Design of Fronthauling and Hybrid Beamforming for Downlink C-RAN Systems," in IEEE Transactions on Communications, vol. 67, no. 6, June 2019.

[16] M. Moltafet, S. Parsaeefard, M. R. Javan and N. Mokari, "Robust Radio Resource Allocation in MISO-SCMA Assisted C-RAN in 5G Networks," in IEEE Transactions on Vehicular Technology, vol. 68, no. 6, pp. 5758-5768, June 2019.

[17] W. Tang and S. Feng, "User Selection and Power Minimization in Full-Duplex Cloud Radio Access Networks," in IEEE Transactions on Signal Processing, vol. 67, no. 9, pp. 2426-2438, 1 May1, 2019.

[18] Y. Wu, A. Deligiannis and S. Lambotharan, "Coalitional Games for Downlink Multicell Beamforming," in IEEE Access, vol. 5, pp. 9251-9265, 2017.

[19] C. Zhang, M. Dong and K. Ota, "Fine-Grained Management in 5G: DQL Based Intelligent Resource Allocation for Network Function Virtualization in C-RAN," in IEEE Transactions on Cognitive Communications and Networking, vol. 6, no. 2, pp. 428-435, June 2020.

[20] N. Mharsi and M. Hadji,"Edge computing optimization for efficient RRH-BBU assignment in cloud radio access networks," in computer networks Journal, Vol. 164, 9 December 2019.

[21] J. Riihijarvi, M. Petrova and P. Mahonen, "Frequency allocation for WLANs using graph colouring techniques," Second Annual Conference on Wireless On-demand Network Systems and Services, St. Moritz, Switzerland, 2005, pp. 216-222.

[22] S. Noh, M. D. Zoltowski, Y. Sung and D. J. Love, "Pilot Beam Pattern Design for Channel Estimation in Massive MIMO Systems," in IEEE Journal of Selected Topics in Signal Processing, vol. 8, no. 5, Oct. 2014.

[23] G. Auer et al., "How much energy is needed to run a wireless network?" IEEE Trans. Wireless Commun., vol. 18, no. 5, pp. 40–49, Oct. 2011.

[24] J. Gao, S. A. Vorobyov, H. Jiang, J. Zhang and M. Haardt, "Sum-Rate Maximization With Minimum Power Consumption for MIMO DF Two-Way Relaying— Part II: Network Optimization," in IEEE Transactions on Signal Processing, vol. 61, no. 14, pp. 3578-3591, July15, 2013.

[25] H. Tuy, M. Minoux, and N. Hoai-Phuong, "Discrete monotonic optimization with application to a discrete location problem," SIAM Journal on Optimization, vol. 17, no. 1, pp. 78–97, 2006.

[26] G. Scutari, F. Facchinei, L. Lampariello and P. Song, "Parallel and distributed methods for nonconvex optimization-Part III: Theory Applications," IEEE Trans. on Signal Processing, pp. 840-844, May.2014.

[27] Kuri-Morales A.F., Gutiérrez-García J. Penalty Function Methods for Constrained Optimization with Genetic Algorithms: A Statistical Analysis., Lecture Notes in Computer Science, Springer, 2002.'

[28] D. T. Ngo, S. Khakurel and T. Le-Ngoc, "Joint Subchannel Assignment and Power Allocation for OFDMA Femtocell Networks," in IEEE Transactions on Wireless Communications, vol. 13, January 2014.

[29] Ferdosian, N., Othman, M., Ali, B.M. et al. Greedy–knapsack algorithm for optimal downlink resource allocation in LTE networks. Wireless Netw 22, 1427–1440 (2016).

[30] Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: Simplified Data Processing on Large Clusters. Commun. ACM 51 (2008), 107–113.

Payam Rahimi (S'19) received the BSc (2010) and MSc (2015) degrees in computer engineering from the Qazvin Azad University, Qazvin, Iran. He is currently working toward the Ph.D. degree in computer engineering at Frederick University, Nicosia, Cyprus, where he is also an active researcher with the Networks Research Laboratory, Department of Electrical Engineering, Computer Engineering and Informatics, Frederick University. His research interests include 5G and beyond network optimization, radio resource allocation, software-defined networking, virtualization, and fog computing.

Chrysostomos Chrysostomou (S'04, M'07) is currently an Assistant Professor with the Department of Electrical Engineering, Computer Engineering and Informatics, at Frederick University, Cyprus, and the Director of the Networks Research Laboratory. His research interests include Quality of Service provisioning in mobile/wireless networks, including Internet-of-Things (IoT) architectures, Device-to-Device (D2D) Communications in 5G Networks for energy and spectrum efficiency, cooperative multi-hop 5G D2D communication with energy harvesting capabilities for public safety networks. Moreover, current focus is given to multi-objective optimization of radio resource allocation in software defined networking based virtualized fog computing radio access networks in 5G-and-beyond wireless environments, mobility management in maritime IoT applications, blockchain technology in communication networks. In addition, research work has been conducted in the mobility support in wireless sensor networks, and in intelligent, QoS-aware mechanisms in vehicular ad hoc networks.

Haris Pervaiz (S'09, M'09) is an assistant professor with the School of Computing and Communications (SCC), Lancaster University, U.K. From April 2017 to October 2018, he was a research fellow with the 5G Innovation Centre, University of Surrey, U.K. From 2016 to 2017, he was an EPSRC Doctoral Prize Fellow with the SCC, Lancaster University. He received his Ph.D. degree from Lancaster University, U.K., in 2016. His current research interests include green heterogeneous wireless communications and networking, 5G and beyond, millimeter wave communication, and energy and spectral efficiency.

Vasos Vassiliou (M'96, SM'19) is an Associate Professor at the Computer Science Department of the University of Cyprus and the co-Director of the Networks Research Laboratory. He is also the Group Leader of the Smart Networked Systems Research Group of the CYENS Center of Excellence, situated in Nicosia, Cyprus. He has published more than 80 technical articles and scientific papers on topics that include Next Generation Network Architectures, Mobile Protocols, Mobile Networks, Wireless Communications and QoS and Traffic Engineering for computer and telecommunication networks. Dr. Vassiliou's current research work is in the 5G and 6G Networks and the Internet of Things domains, where he and his team work on Security, Privacy, Mobility, Resource Management, and Data Management issues.

Qiang Ni (M'04, SM'08) is a Professor at the School of Computing and Communications, Lancaster University, Lancaster, U.K. His research interests include the area of future generation communications and networking, including green communications and networking, millimeter-wave wireless communications, cognitive radio network systems, non-orthogonal multiple access (NOMA), heterogeneous networks, 5G and 6G, SDN, cloud networks, edge computing, dispersed computing, energy harvesting, wireless information and power transfer, IoTs, cyber physical systems, AI and machine learning, big data analytics, and vehicular networks. He has authored or co-authored 300+ papers in these areas. He was an IEEE 802.11 Wireless Standard Working Group Voting Member and a contributor to various IEEE wireless standards.