Computation Offloading for IoT in C-RAN: Optimization and Deep Learning

Chandan Pradhan, *Student Member, IEEE*, Ang Li, *Member, IEEE*, Changyang She, *Member, IEEE*, Yonghui Li, *Fellow, IEEE*, and Branka Vucetic, *Life Fellow, IEEE*

***Abstract*— We consider computation-offloading for Internet-of-things (IoT) applications in multiple-input-multiple-output (MIMO) cloud-radio-access-network (C-RAN). Specifically, the computational tasks of the IoT devices (IoTDs) are offloaded to a MIMO C-RAN, where a MIMO radio resource head (RRH) is connected to a baseband unit (BBU) through a capacity-limited fronthaul link, facilitated by the spatial filtering and uniform scalar quantization. We formulate a computation-offloading optimization problem to minimize the total transmit power of the IoTDs while satisfying the latency requirement of the computational tasks. To obtain a feasible solution for the non-convex problem, firstly the spatial filtering matrix is locally optimized at the MIMO RRH. Subsequently, leveraging the alternating optimization framework for joint optimization on the residual variables at the BBU, the baseband combiner, the optimal resource allocation and the number of quantization bits are obtained through the minimum-mean-squared-error (MMSE) metric, the successive inner convexification method and the line-search method, respectively. As a low-complexity approach, we apply a supervised deep learning (DL) method, which learns from the solutions obtained with our proposed algorithm. In addition, the deep transfer learning is adopted to adjust the neural network in dynamic IoT systems. Numerical results validate the effectiveness of the proposed optimization algorithm and the learning based methods.**

***Index Terms*— IoT, computation-offloading, C-RAN, deep learning, deep transfer learning.**

I. Introduction

INTERNET-OF-THINGS (IoT) has a great potential to impact our lives in the future by providing solutions related to multiple sectors of industry, smart homes, transportation, etc. It is predicted that there will be about 50 billion IoT devices by 2020 [1], [2]. The deployment of a large-scale IoT ecosystem requires the IoT devices (IoTDs) with a small physical size to be built from cost-efficient hardware components, which results in major challenges due to their limited battery life and computational capability. More importantly, IoT applications require flexibility in handling diverse latency requirements [1], [3]. To address the limited battery life and computational capability, the computational task in an IoTD can be migrated to a more powerful server [4], which is known as *computation-offloading* [4], [5]. Furthermore, technologies like massive multiple-input-multiple-output (MIMO) [6]–[8] and cloud-radio-access-network (C-RAN) [9]–[15] can be exploited to augment the process of computation-offloading and manage the corresponding latency requirement imposed by the IoT applications.

Massive MIMO, characterized by the deployment of a huge number of antennas, is a key enabling technique for 5G wireless systems [6], [7]. More recently, *extra-large scale MIMO (xL-MIMO)* as a step further has received increasing research attention [8], [16]–[18]. In xL-MIMO, a large antenna array in the order of hundreds and thousands is integrated into a large man-made structure, for example, walls of buildings in the residential rooms, airports, or large shopping malls, as a scaled-up version of the massive MIMO systems where the spatial dimension provides an additional degree of freedom to further enhance the performance of the massive MIMO systems [16], [19]. In addition, xL-MIMO systems provide a better coverage with a line-of-sight (LOS) channel, which also simplifies the corresponding channel estimation [8], [17], [18]. However, the implementation of such xL-MIMO systems is challenging due to the deployment complexity along with the increasing requirement for the baseband signal processing, which is proportional to the number of antenna elements.

C-RAN can be a potential technique to overcome the above challenges for the xL-MIMO systems. Specifically, C-RAN migrates the baseband signal processing to a baseband unit (BBU) that is equipped with a powerful server in the "cloud", while the radio frequency (RF) functionalities are implemented at the remote radio head (RRH) [20]. By combining massive MIMO with C-RAN, the deployment complexity of the conventional massive MIMO systems can be greatly alleviated, since only analog components such as

Manuscript received September 9, 2019; revised January 29, 2020; accepted March 17, 2020. Date of publication March 25, 2020; date of current version July 15, 2020. This work was supported in part by the Science and Technology Program of Shaanxi Province No. 2019KW-007, in part by the Australian Research Council (ARC) under Grant DP190101988, and in part by the ARC Laureate Fellowship under Grant FL160100032. The associate editor coordinating the review of this article and approving it for publication was L. Wei. *(Corresponding author: Ang Li.)*

Chandan Pradhan, Changyang She, Yonghui Li, and Branka Vucetic are with the Centre of Excellence in Telecommunications, School of Electrical and Information Engineering, The University of Sydney, Sydney, NSW 2006, Australia (e-mail: chandan.pradhan@sydney.edu.au; changyang.she@sydney.edu.au; yonghui.li@sydney.edu.au; branka.vucetic@sydney.edu.au).

Ang Li was with the School of Electrical and Information Engineering, The University of Sydney, Sydney, NSW 2006, Australia. He is now with the Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: ang.li.2020@xjtu.edu.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCOMM.2020.2983142

0090-6778 © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on April 29,2022 at 09:05:06 UTC from IEEE Xplore. Restrictions apply.

antennas and RF chains with a limited signal processing capability are required [9], [13]. However, moving the signal processing of a massive MIMO system from the RRH to the central BBU requires a huge amount of digitally sampled data to be transmitted over the fronthaul link. Therefore, it is necessary to compress the uplink data at the RRH to satisfy the capacity limit of the fronthaul link. Accordingly, in [9], the authors proposed a data compression method which reduces the dimension of the signals received across the multiple antennas through spatial filtering, followed by a uniform scalar quantization across the reduced dimension. To further reduce the cost and power consumption of the hardware components in a C-RAN system, hybrid analog-digital designs have subsequently been applied to the massive MIMO C-RANs [12], [13], [21]–[24], where the number of RF chains at the RRH can be reduced.

Different from the wireless systems in [9] and [13] where the uplink communication has a high spectral efficiency requirement, the latency-constrained IoT applications pursue low data rates with a higher energy efficiency performance while meeting their stringent latency constraints. In this regard, computation-offloading with the massive MIMO C-RAN can be leveraged by allocating the transmit power and computational resource at the BBU server to each IoTD while satisfying their latency requirement. While there have been recent studies on the joint admission control and computation-offloading strategies [25], [26], the joint computation-offloading and data caching strategies in a hybrid mobile cloud/edge computation system [27], and the joint optimization of the trajectory, task data and computing resource allocations in a non-orthogonal multiple access (NOMA)-based and unmanned aerial vehicle (UAV)-assisted edge computing system [28] for a single-antenna wireless network, there are only a limited number of works in the literature that study the joint communication and computational resource allocation for computation-offloading in a MIMO C-RAN [10], [11], [14]. In [10] and [11], the offloading problems were formulated to minimize the total transmit power and energy consumption of the devices, respectively, while meeting the latency constraints. The computation-offloading method proposed in [14] aimed to minimize the maximum latency of all the devices. Nevertheless, these works did not consider the compression and quantization of the received signal at the RRH as in [9], [13], which can lead to infeasible data traffic for a capacity-limited fronthaul link. Moreover, the latency incurred at the capacity-limited fronthaul link in transferring the data from the RRH to the BBU can critically impede the execution of the computational tasks. Accordingly, the above mentioned drawbacks call for the joint design of the resource allocations, the compression and quantization strategies, especially for the latency-critical IoT applications.

Furthermore, the iterative nature of the solutions proposed in [9]–[11], [13], [14] are computationally demanding for real-time implementation, especially when the computational tasks have a stringent latency requirement. Recently, deep learning (DL) has become a promising tool in approximating the optimal policies for wireless communication problems [29]–[33], which can return a near-optimal solution with a simple operation such as matrix-vector multiplication. Specifically, based on the universal approximation theorem [34], DL approximates the policy of a given complicated optimization problem by using a *deep neural network* (DNN), where the DNN can be trained offline with the label samples given by the outputs obtained from running a computationally expensive algorithm corresponding to the optimization problem [35], [36]. Moreover, considering that the dynamic operating parameters of the MIMO C-RAN network can perturb the optimal policy of the underlying optimization problem, deep transfer learning (DTL) can be leveraged to fine-tune a pre-trained DNN to approximate the updated optimal policy with considerably lesser training data samples [37], [38].

Motivated by the above, in this work we consider the computation-offloading problem for the IoTDs in a massive MIMO C-RAN deployed in an indoor environment, where multiple receive antennas that are spread across one of the walls act as an xL-MIMO RRH. Specifically, the uplink signals, encoding the computation bits from the IoTDs, are firstly received at the xL-MIMO before spatial filtering. Subsequently, the filtered signals are quantized [39] and transmitted to the BBU via a capacity-limited fronthaul link, where a baseband combiner corresponding to each IoTD extracts and forwards the respective signal to the BBU server. We focus on the minimization of the total transmit power of the IoTDs, while satisfying the latency requirement of their corresponding computational task. We summarize the main contributions of the paper below:

1) We establish a computation-offloading optimization problem to minimize the total transmit power of the IoTDs by jointly optimizing the spatial filtering design at the xL-MIMO RRH, the communication and computational resource allocation policy, the number of quantization bits, and the baseband combiner design at the BBU, while satisfying the latency requirement of the corresponding computational tasks. Compared to [10] and [11] where the latency requirement only includes the transmission latency and computational latency, we further consider the fronthaul latency experienced in transferring the quantized bits from the xL-MIMO RRH to the BBU. This additional fronthaul latency couples with the transmission latency through the required number of quantization bits and makes our non-convex optimization problem fundamentally different from the existing works [10], [11], [14], which is more challenging to solve.
2) To obtain a near-optimal solution for the formulated optimization problem, we introduce a two-stage design, where a hybrid spatial filtering (HSF) matrix at the xL-MIMO RRH is firstly obtained purely based on the channel state information (CSI). Subsequently, based on the effective channel and the obtained HSF matrix, a joint optimization on the residual variables at the BBU is implemented. For the joint optimization at the BBU, the proposed problem is divided into three sub-problems and solved via alternating optimization. To be more

Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on April 29,2022 at 09:05:06 UTC from IEEE Xplore. Restrictions apply.

specific, the baseband combiner is obtained following the minimum-mean-squared-error (MMSE) metric, the communication and computational resource allocation problem is solved by leveraging the successive inner convexification, and the optimization on the number of quantization bits is solved through a line-search method. Moreover, the proposed algorithm is shown to converge to a local optimal solution.

3) For practical implementation, we resort to DL as a low-complexity solution to the joint optimization at the BBU [30]–[33], [40]. Specifically, we deploy a supervised learning method using the DNN, where the solutions obtained from the complicated optimization algorithm are used as labeled training samples. To further enhance the adaptability of the DNN to the dynamic operating parameters, we have adopted DTL to fine-tune the pre-trained DNN with a small number of new training data samples [37], [38]. Finally, the numerical results demonstrate the superiority of the proposed joint optimization algorithm over the disjoint optimization procedures. Furthermore, the DNN based supervised learning along with the DTL is shown to be an effective low-complexity approach, which reduces the execution time by the order of two magnitudes.

The rest of the paper is organized as follows. Section II describes the system model and introduces the formulated problem for the total transmit power minimization. In Section III, we present the proposed solution for the formulated problem, where the HSF matrix at the xL-MIMO RRH is obtained locally, followed by the joint optimization of the residual parameters at the BBU via alternating optimization. Section IV discusses the low-complexity solution for the joint optimization based on the DNN, where the adaptability of the DNN is enhanced through DTL. Numerical results are presented in Section V, and we conclude the paper in Section VI.

Notations: Bold upper-case letters $\mathbf{Y}$, bold lower-case letters $\mathbf{y}$ and letters y denote matrices, vectors and scalars, respectively; $Y_{i,j}$ is the entry on the i-th row and j-th column of $\mathbf{Y}$; Transpose and conjugate transpose of $\mathbf{Y}$ are represented by $\mathbf{Y}^T$ and $\mathbf{Y}^H$, respectively; $\mathbf{Y}^\dagger$ is the Moore-Penrose pseudo inverse of $\mathbf{Y}$; $\text{diag}\left([y_1,\ldots,y_n]^T\right)$ denotes a diagonal matrix with elements y_i, $i=1,\ldots,n$ on the diagonal; $\text{vec}(\mathbf{Y})$ indicates vectorization; $\|\mathbf{y}\|_2$ is the ℓ_2 norm of the vector $\mathbf{y}$; $\mathbf{1}_M$ is the $M\times 1$ vector of ones; $\jmath$ is defined as $\jmath \triangleq \sqrt{-1}$, $|\cdot|$ returns the amplitude of a complex number; $\odot$, $\oslash$ and $\circ$ denote the Hadamard product, division and power, respectively; $\mathbf{I}_m \in \mathbb{R}^{m\times m}$ is the identity matrix.

II. System Model and Problem Formulation

We consider the uplink of an xL-MIMO C-RAN that serves K single-antenna IoTDs, as shown in Fig. 1. The xL-MIMO RRH consists of N antennas uniformly distributed in a two-dimensional space along the xz-plane at $y=0$ on the Cartesian coordinates. Accordingly, the locations of the n-th antenna of the xL-MIMO RRH and the k-th IoTD are defined as $(\bar{x}_n, 0, \bar{z}_n)$ and (x_k, y_k, z_k), respectively. In this

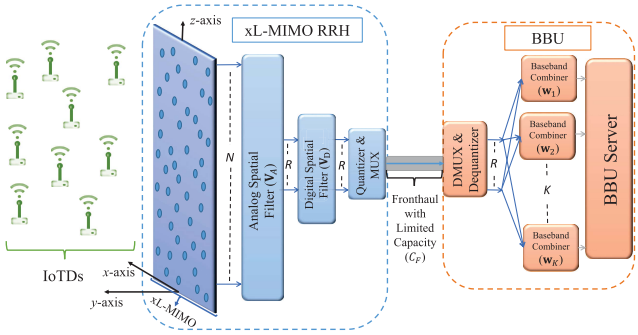


Fig. 1. System model of the uplink xL-MIMO C-RAN serving K IoTDS.

paper, we assume that the xL-MIMO RRH is equipped with $R=K\,(\ll N)$ RF chains such that there are enough spatial degrees of freedom to serve all the K IoTDs [13]. The xL-MIMO RRH is connected to the BBU[1] via a digital error-free fronthaul link with a capacity of C_F bits per second (bps). The BBU makes the resource allocation decisions and decodes the IoTDs' symbols, followed by the processing of the computation bits at the BBU server. We assume that all the IoTDs transmit over a quasi-static flat-fading channel, and the received signal at the xL-MIMO RRH is expressed as

$$\mathbf{y}=\sum_{k=1}^{K}\mathbf{h}_k\sqrt{p_k}s_k+\mathbf{z}, \tag{1}$$

where s_k is the transmitted symbol of the k-th IoTD such that $|s_k|^2=1$, p_k is the corresponding transmit power, $\mathbf{h}_k\in\mathbb{C}^{N\times 1}$ is the channel vector between the k-th IoTD and the xL-MIMO RRH, and $\mathbf{z}\sim\mathcal{CN}(0,\sigma^2\mathbf{I}_N)$ denotes the additive white Gaussian noise (AWGN).

A. Channel Model

Given that the IoTDs are deployed in an indoor environment, the IoTDs are reasonably close to the xL-MIMO RRH. Hence, the desired channel between an IoTD and each antenna of the xL-MIMO RRH is composed of both the deterministic LOS and non-line-of-sight (NLOS) components. Accordingly, the channel between the k-th IoTD and the xL-MIMO RRH is given by [8]

$$\mathbf{h}_k=\boldsymbol{\kappa}_k^L\mathbf{h}_k^L+\boldsymbol{\kappa}_k^{NL}\mathbf{h}_k^{NL}, \tag{2}$$

where $\mathbf{h}_k^L\in\mathbb{C}^{N\times 1}$ is the deterministic LOS component between the k-th IoTD and the xL-MIMO RRH, given by [8]

$$\mathbf{h}_k^L=\left[l_{1,k}^L h_{1,k},\ldots,l_{N,k}^L h_{N,k}\right]^T, \tag{3}$$

where $l_{n,k}^L=\frac{1}{\sqrt{4\pi d_{n,k}^2}}$, $h_{n,k}=\exp\left(\frac{-j2\pi d_{n,k}}{\lambda}\right)$ and $d_{n,k}=\sqrt{(x_k-\bar{x}_n)^2+y_k^2+(z_k-\bar{z}_n)^2}$ are the attenuation factor in the free space, the channel gain and the distance between the k-th IoTD and the n-th antenna of the xL-MIMO RRH, respectively, with λ denoting the carrier wavelength of the transmitted signal. The NLOS component $\mathbf{h}_k^{NL}\in\mathbb{C}^{N\times 1}$

[1]Note that the BBU can be shared among multiple xL-MIMO systems, thereby reducing the total cost of ownership (TCO) [20].

Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on April 29,2022 at 09:05:06 UTC from IEEE Xplore. Restrictions apply.

between the k-th IoTD and the xL-MIMO RRH is defined as [32]

$$\mathbf{h}_k^{NL} = \mathbf{\Lambda}_k^{\frac{1}{2}} \mathbf{g}_k, \tag{4}$$

with $\mathbf{\Lambda}_k \triangleq \text{diag}\left(\left[d_{1,k}^{-\delta}\tau_{1,k}, \ldots, d_{N,k}^{-\delta}\tau_{N,k}\right]^T\right) \in \mathbb{C}^{N\times N}$, where $d_{n,k}^{-\delta}$ and $\tau_{n,k}$ are the large-scale fading and the log-normal shadow fading between the k-th IoTD and the n-th antenna of the xL-MIMO RRH, respectively. Furthermore, $10\log_{10}(\tau_{n,k}), \forall n,k$ is a zero-mean Gaussian random variable with a standard deviation of σ_{shad}, which is assumed to change slowly with time and known a priori [6], [41]. δ is the path loss exponent and $\mathbf{g}_k \in \mathbb{C}^{N\times 1}$ models the small-scale fading, with each entry following $\mathcal{CN}(0,1)$. Finally, $\boldsymbol{\kappa}_k^L \triangleq \text{diag}\left(\left[\sqrt{\frac{\kappa_{1,k}}{\kappa_{1,k}+1}}, \ldots, \sqrt{\frac{\kappa_{N,k}}{\kappa_{N,k}+1}}\right]^T\right) \in \mathbb{R}^{N\times N}$ and $\boldsymbol{\kappa}_k^{NL} \triangleq \text{diag}\left(\left[\sqrt{\frac{1}{\kappa_{1,k}+1}}, \ldots, \sqrt{\frac{1}{\kappa_{N,k}+1}}\right]^T\right) \in \mathbb{R}^{N\times N}$, where $\kappa_{n,k}$ denotes the Rician factor between the k-th IoTD and the n-th antenna of the xL-MIMO RRH.

B. Uplink Signal Processing

At the xL-MIMO RRH, we consider the spatial-compression-and-forward (SCF) scheme proposed in [9], [13], [39] to balance between the information conveyed to the BBU and the data traffic over the fronthaul link. To reduce the hardware complexity, we employ the hybrid analog-digital spatial filtering, where each antenna is only equipped with a phase shifter and the signals from N antennas are filtered using an analog spatial filtering matrix $\mathbf{V}_A \in \mathbb{C}^{N\times R}$, followed by a digital spatial filtering matrix $\mathbf{V}_D \in \mathbb{C}^{R\times R}$. Accordingly, the received signal after the HSF is given by

$$\bar{\mathbf{y}} = \mathbf{V}\mathbf{y} = \mathbf{V}\sum_{k=1}^{K} \mathbf{h}_k \sqrt{p_k} s_k + \mathbf{V}\mathbf{z}, \tag{5}$$

where $\mathbf{V} \triangleq \mathbf{V}_D^H \mathbf{V}_A^H = [\mathbf{v}_1, \ldots, \mathbf{v}_R]^T \in \mathbb{C}^{R\times N}$ denotes the HSF matrix. With the use of phase shifters, each entry of $\mathbf{V}_A$ satisfies the element-wise constant-modulus constraint, i.e., $\left|[\mathbf{V}_A]_{(i,j)}\right| = 1, \ \forall i,j$. In this paper, we assume that high-resolution ADCs are used at the xL-MIMO RRH such that the quantization error due to ADCs is negligible [13]. Subsequently, a uniform scalar quantization is applied to each element of $\bar{\mathbf{y}} = [\bar{y}_1, \ldots, \bar{y}_R]^T$, where each complex symbol $\bar{y}_r$ can be represented by its in-phase (I) and quadrature (Q) part as

$$\bar{y}_r = \bar{y}_r^I + \jmath\, \bar{y}_r^Q, \quad \forall r, \tag{6}$$

where the I-branch symbol $\bar{y}_r^I$ and Q-branch symbol $\bar{y}_r^Q$ are both real Gaussian random variables with zero mean and variance $\left(\sum_{k=1}^{K} p_k |\mathbf{v}_r^T \mathbf{h}_k|^2 + \sigma^2 \parallel \mathbf{v}_r \parallel^2\right)/2$ [39]. After the uniform scalar quantization, the baseband quantized symbol of $\bar{\mathbf{y}}$ is given by

$$\tilde{\mathbf{y}} = \bar{\mathbf{y}} + \mathbf{e} = \mathbf{V}\sum_{k=1}^{K} \mathbf{h}_k \sqrt{p_k} s_k + \mathbf{V}\mathbf{z} + \mathbf{e}, \tag{7}$$

where $\mathbf{e} \triangleq [e_1, \ldots, e_R]^T$ denotes the additive quantization error vector for $\bar{\mathbf{y}}$. Each e_r is Gaussian distributed with zero mean and variance ϱ_r, with ϱ_r given by [39]

$$\varrho_r = \begin{cases} 3\left(\sum_{k=1}^{K} p_k \left|\mathbf{v}_r^T \mathbf{h}_k\right|^2 + \sigma^2 ||\mathbf{v}_r||^2\right) 2^{-2\varpi}, & \text{if } \varpi > 0, \\ \infty, & \text{if } \varpi = 0, \end{cases} \tag{8}$$

where ϖ denotes the number of bits that each RF chain uses to quantize $\bar{y}_r^I$ and $\bar{y}_r^Q$. As each e_r is independent over r due to the independent scalar quantization for each element of $\bar{\mathbf{y}}$, and therefore the covariance matrix of $\mathbf{e}$ is a function of $\mathbf{p} \triangleq [p_1, \ldots, p_K]^T$, $\mathbf{V}$ and ϖ, given by

$$\mathbf{Q}(\mathbf{p}, \mathbf{V}, \varpi) = \mathbb{E}[\mathbf{e}\mathbf{e}^H] = \text{diag}\left([\varrho_1, \ldots, \varrho_r]^T\right). \tag{9}$$

Subsequently, the quantized symbols are forwarded to the BBU via the fronthaul link. To mitigate the effects of the inter-IoTD interference and the quantization error, a linear baseband combiner $\mathbf{w}_k \triangleq [w_{k,1}, \ldots, w_{K,R}]^T \in \mathbb{C}^{R\times 1}$ is further applied to $\tilde{\mathbf{y}}$ before demodulating the symbol for the k-th IoTD, given by

$$\begin{aligned} \hat{s}_k &= \mathbf{w}_k^H \tilde{\mathbf{y}}, \\ &= \mathbf{w}_k^H \mathbf{V}\mathbf{h}_k \sqrt{p_k} s_k + \sum_{j=1,\, j\neq k}^{K} \mathbf{w}_k^H \mathbf{V}\mathbf{h}_j \sqrt{p_j} s_j + \mathbf{w}_k^H \mathbf{V}\mathbf{z} \\ &\quad + \mathbf{w}_k^H \mathbf{e}. \end{aligned} \tag{10}$$

Accordingly, the SINR for the k-th IoTD is expressed as

$$\begin{aligned} &\gamma_k(\mathbf{p}, \mathbf{V}, \mathbf{W}, \varpi) \\ &= \frac{p_k \left|\mathbf{w}_k^H \mathbf{V}\mathbf{h}_k\right|^2}{\sum_{j=1, j\neq k}^{K} p_j \left|\mathbf{w}_k^H \mathbf{V}\mathbf{h}_j\right|^2 + \sigma^2 \| \mathbf{w}_k^H \mathbf{V}\|^2 + \mathbf{w}_k^H \mathbf{Q}(\mathbf{p}, \mathbf{V}, \varpi)\mathbf{w}_k}, \end{aligned} \tag{11}$$

where $\mathbf{W} \triangleq [\mathbf{w}_1, \ldots, \mathbf{w}_K]$.

C. Computation-Offloading and Latency Model

We assume that due to the limited computational capability at the IoTDs, all the computational tasks of the IoTDs have to be offloaded to the BBU. Accordingly, let the k-th IoTD's computational task C_k be described by a tuple, defined as $\left(\omega_k, b_k, \mathcal{T}_k^{th}\right)$, where ω_k denotes the number of CPU cycles needed for computing C_k, b_k represents the number of computation bits needed for C_k and $\mathcal{T}_k^{th}$ is the maximum tolerable latency to execute C_k [11]. In the case of offloading, the latency includes a) the transmission latency, b) the fronthaul latency, and c) the computational latency.

1) Transmission Latency (ξ_k^{TL}): The transmission latency ξ_k^{TL} is incurred during the transmission of the computation bits b_k from the k-th IoTD to the xL-MIMO RRH. Accordingly, given $\gamma_k(\mathbf{p}, \mathbf{V}, \mathbf{W}, \varpi), \ \forall k$, the latency for the transmission of $\log_2(1 + \gamma_k(\mathbf{p}, \mathbf{V}, \mathbf{W}, \varpi))$ bits per second per Hertz is given by [10], [11], [14]

$$\xi_k^{TL} = \frac{b_k}{B_W \log_2(1 + \gamma_k(\mathbf{p}, \mathbf{V}, \mathbf{W}, \varpi))}, \tag{12}$$

where B_W is the total transmission bandwidth.

Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on April 29,2022 at 09:05:06 UTC from IEEE Xplore. Restrictions apply.

2) Fronthaul Latency (ξ_k^{FL}): For b_k computation bits corresponding to the k-th IoTD, we assume that the bits are encoded using the M-PSK modulation. Accordingly, b_k bits are encoded into $\frac{b_k}{\log_2(M)}$ symbols which are transmitted from the k-th IoTD to the BBU through the xL-MIMO RRH. With ϖ bits used to quantize both the real and imaginary parts of each entry in $\bar{\mathbf{y}}$, a total of $2R\varpi$ quantized bits are required across R RF chains [9], [39]. Consequently, $\frac{b_k}{\log_2(M)}$ transmitted symbols of the k-th IoTD generate an effective traffic of $\frac{2b_kR\varpi}{\log_2(M)}$ bits for the fronthaul link. Hence, with a fronthaul link capacity of C_F, expressed in terms of bits per second, the fronthaul latency ξ_k^{FL} for forwarding b_k computation bits of the k-th IoTD from the xL-MIMO RRH to the BBU is given by [14]

$$\xi_k^{FL} = \frac{2b_kR\varpi}{C_F \log_2(M)}. \tag{13}$$

3) Computational Latency (ξ_k^{CL}): The computational resources are shared among the K IoTDs and are quantified by the computational rate F_T, expressed in terms of the number of CPU cycles per second [11], [14]. Let us denote by $f_k \geq 0$ the fraction of F_T to be assigned to each IoTD. The rates f_k are subject to the computational budget constraint, i.e.,

$$\sum_{k=1}^{K} f_k \leq F_T. \tag{14}$$

Given the resource assignment f_k, the computational latency ξ_k^{CL} incurred in executing ω_k CPU cycles for the computational task of the k-th IoTD is given by [10], [11], [14]

$$\xi_k^{CL} = \frac{\omega_k}{f_k}, \quad \forall k. \tag{15}$$

Finally, the expression for the overall latency ξ_k is given by

$$\begin{aligned}\xi_k &= \xi_k^{TL} + \xi_k^{FL} + \xi_k^{CL}, \\ &= \frac{b_k}{B_W \log_2\left(1+\gamma_k\left(\mathbf{p},\mathbf{V},\mathbf{W},\varpi\right)\right)} + \frac{2b_kR\varpi}{C_F\log_2(M)} + \frac{\omega_k}{f_k}.\end{aligned} \tag{16}$$

(16) clearly shows the interplay between the wireless transmission part and the computational part via the transmission and computational latency. Furthermore, a coupling between the transmission and fronthaul latency through the number of quantization bits ϖ can also be observed from (16). For example, an increase in the quantization bits decreases the quantization error which reduces the transmission latency, while on the other hand, it increases the required number of bits transmitted to the BBU, thereby increasing the fronthaul latency. Therefore, the joint optimization of the communication and computational resource allocations along with the number of quantization bits for the computation-offloading task is essential.

D. Problem Formulation

In this paper, we aim to minimize the total transmit power for the IoTDs, i.e., $\mathbf{1}_K^T\mathbf{p}$, by jointly optimizing the HSF matrix $\mathbf{V}$, the baseband combiner $\mathbf{W}$, the communication resource $\mathbf{p}$, the computational resource $\mathbf{f} \triangleq [f_1,\ldots,f_K]^T$ and the number of quantization bits ϖ. Accordingly, we aim to solve the following optimization problem:

$$\begin{aligned}\mathcal{P}_1: \ &\min_{\{\mathbf{V},\mathbf{W},\mathbf{p},\mathbf{f},\varpi\}} \mathbf{1}_K^T\mathbf{p} \\ &\textit{s.t.}\ \mathrm{C}_1: \xi_k \leq \mathcal{T}_k^{th}, \quad \forall k, \\ &\quad \mathrm{C}_2: \mathbf{1}_K^T\mathbf{f} \leq F_T, \quad \mathrm{C}_3: p_k \leq P_{k,max}, \quad \forall k, \\ &\quad \mathrm{C}_4: 2B_WR\varpi \leq C_F, \quad \mathrm{C}_5: \varpi \in \mathbb{Z}_{>0}, \\ &\quad \mathrm{C}_6: \left|[\mathbf{V}_A]_{(i,j)}\right| = 1, \quad \forall i,j,\end{aligned} \tag{17}$$

where C_1 is the latency constraint for the k-th IoTD with the latency threshold denoted by $\mathcal{T}_k^{th}$, C_2 is the computational resource constraint, C_3 is the maximum power limits of the IoTDs, where $P_{k,max}$ denotes the maximum transmit power of each IoTD, C_4 is the fronthaul capacity constraint [9], [11], [13], C_5 is the integer constraint for the number of quantization bits, and C_6 is the element-wise constant-modulus constraint for the analog spatial filtering matrix.

III. Proposed Solution for the Formulated Problem $\mathcal{P}_1$

In this section, we seek a feasible solution for $\mathcal{P}_1$, which is found to be non-convex due to 1) the coupling of variables between the transmission latency, the fronthaul latency and the computational latency, 2) the integer constraint for the quantization bit, and 3) the element-wise constant-modulus constraint for the analog spatial filtering matrix. Accordingly, to solve $\mathcal{P}_1$, we present a two-stage design, where the HSF matrix at the xL-MIMO RRH is obtained locally[2] based on the CSI, and a joint optimization on the residual variables at the BBU is subsequently implemented based on the effective channel and the obtained HSF matrix.

A. HSF Design at the xL-MIMO RRH

In this work, the HSF matrix is obtained by approximating the fully-digital spatial filtering (FDSF) matrix. To pursue a low-complexity solution, we select the matched filtering (MF) method as the FDSF, given by

$$\mathbf{V}_{FD} = \mathbf{H}^H, \tag{18}$$

where $\mathbf{H} \triangleq [\mathbf{h}_1,\ldots,\mathbf{h}_K]$. Note that MF leads to the scaling down of the optimal transmit power of each IoTDs with an increase in N as analysed in [6], [41]. Another advantage of employing the MF approach is that $\mathbf{V}_{FD}$ tends to eliminate the effect of small-scale fading, resulting in a less frequent update of the communication and computational parameters at the BBU. Specifically, assuming the independence of propagation paths of the IoTDs and by leveraging the concepts of channel hardening and law of large numbers [6], [8], it can be shown that the effective channel at the BBU, given by $\mathbf{V}_{FD}\mathbf{H}$, tends to be independent of the small-scale fading, i.e.,

$$\mathbf{V}_{FD}\mathbf{H} \xrightarrow{N\to\infty} \mathrm{diag}\left([h_1^e,\ldots,h_K^e]^T\right), \tag{19}$$

[2]Designing the HSF matrix locally at the xL-MIMO RRH reduces the signaling overhead between the BBU and the xL-MIMO RRH.

Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on April 29,2022 at 09:05:06 UTC from IEEE Xplore. Restrictions apply.

where $h_k^e \triangleq \sum_{n=1}^{N} \frac{1}{\kappa_{n,k}+1}\left(\frac{\kappa_{n,k}}{4\pi d_{n,k}^2} + d_{n,k}^{-\delta} e^{\frac{\sigma_{shad}^2}{2}}\right), \quad \forall k.$ Accordingly, $\mathbf{V}_{FD}$ asymptotically decorrelates the signals from the IoTDs across the K output dimensions. Subsequently, for a given FDSF matrix, the hybrid analog and digital spatial filtering matrix, i.e., $\mathbf{V}_A$ and $\mathbf{V}_D$, are obtained by [42]

$$\mathbf{V}_A = \mathbf{V}_{FD}^H \oslash \left|\mathbf{V}_{FD}^H\right|, \tag{20}$$

and

$$\mathbf{V}_D = \mathbf{V}_A^{\dagger}\mathbf{V}_{FD}^H. \tag{21}$$

Consequently, the HSF matrix is given by $\mathbf{V} = \mathbf{V}_D^H\mathbf{V}_A^H$.

B. Joint Optimization at the BBU

Next, we propose to solve a joint optimization on the residual variables at the BBU based on the alternating optimization framework [43], which effectively removes the coupling between the transmission latency, the fronthaul latency and the computational latency. To be more specific, given the HSF matrix $\mathbf{V}$, $\mathcal{P}_1$ can be transformed into a joint optimization on $\mathbf{W}$, $\mathbf{p}$, $\mathbf{f}$ and ϖ, given by

$$\mathcal{P}_2: \min_{\{\mathbf{W},\mathbf{p},\mathbf{f},\varpi\}} \mathbf{1}_K^T\mathbf{p} \quad \textit{s.t. } \mathrm{C}_1, \mathrm{C}_2, \mathrm{C}_3, \mathrm{C}_4, \mathrm{C}_5. \tag{22}$$

It should be noted that the obtained HSF matrix does not completely decorrelate the signals from the IoTDs due to the finite number of antennas at the xL-MIMO RRH, resulting in the inter-IoTD interference. This along with the quantization noise introduced by the subsequent quantizer may degrade the demodulation performance of the signals at the BBU. Hence, we further adopt a baseband combiner at the BBU to obtain the received symbols as close as possible to the original symbols. Consequently, following the MMSE metric and for a given $\mathbf{p}$, $\mathbf{f}$ and ϖ, the optimal linear baseband combiner for $\mathcal{P}_2$ can be derived as [9], [13]

$$\bar{\mathbf{w}}_k = \sqrt{p_k}\Big[\mathbf{V}\left(\mathbf{H}\,\mathrm{diag}\left(\mathbf{p}\right)\mathbf{H}^H + \sigma^2\mathbf{I}_N\right)\mathbf{V}^H + \mathbf{Q}(\mathbf{p},\mathbf{V},\varpi)\Big]^{-1}\mathbf{V}\mathbf{h}_k. \tag{23}$$

Based on the fact that

$$\bar{\mathbf{w}}_k^H\mathbf{Q}(\mathbf{p},\mathbf{V},\varpi)\bar{\mathbf{w}}_k = \sum_{r=1}^{R}\varrho_r\left|\bar{w}_{k,r}\right|^2 = 2^{-2\varpi}\sum_{r=1}^{R}\Xi_{r,k}\left(\mathbf{p}\right), \tag{24}$$

where $\bar{w}_{i,j}$ denotes the j-th element of $\bar{\mathbf{w}}_i$, $1 \le i \le K$, $1 \le j \le R$ and

$$\Xi_{r,k}(\mathbf{p}) = 3\left|\bar{w}_{k,r}\right|^2\left(\sum_{j=1}^{K} p_j\left|\mathbf{v}_r^T\mathbf{h}_j\right|^2 + \sigma^2||\mathbf{v}_r||^2\right), \quad \forall r, \tag{25}$$

and by defining $\alpha_{k,j} \triangleq \left|\bar{\mathbf{w}}_k^H\mathbf{V}\mathbf{h}_j\right|^2$, $\eta_k \triangleq \sigma^2 \parallel \bar{\mathbf{w}}_k^H\mathbf{V} \parallel^2$, (12) can be expressed as a function of $\mathbf{p}$ and ϖ as

$$\xi_k^{TL}\left(\mathbf{p},\varpi\right) \triangleq \frac{b_k}{B_W\log_2\left(1+\frac{p_k\alpha_{k,k}}{\eta_k+\sum_{j=1,\ j\neq k}^{K} p_j\alpha_{k,j}+2^{-2\varpi}\sum_{r=1}^{R}\Xi_{r,k}(\mathbf{p})}\right)}. \tag{26}$$

Accordingly, $\mathcal{P}_2$ is transformed into a joint optimization on $\mathbf{p}$, $\mathbf{f}$, and ϖ, given by

$$\mathcal{P}_3: \min_{\{\mathbf{p},\mathbf{f},\varpi\}} \mathbf{1}_K^T\mathbf{p} \quad \textit{s.t. } \mathrm{C}_2, \mathrm{C}_3, \mathrm{C}_4, \mathrm{C}_5, \quad \mathrm{C}_7: \xi_k^{TL}\left(\mathbf{p},\varpi\right) + \frac{2b_kR\varpi}{C_F\log_2(M)} + \frac{\omega_k}{f_k} \le \mathcal{T}_k^{th}, \ \forall k. \tag{27}$$

Based on the formulation, we discuss the feasibility of $\mathcal{P}_3$, as shown in Lemma 1 below.

Lemma 1: $\mathcal{P}_3$ admits a non-empty feasible set satisfying all the constraints in (27), if for the given $\mathcal{T}_k^{th} > 0, \quad \forall k, \ \exists \ \mathbf{p} \in \Psi \triangleq \left\{\bar{\mathbf{p}} \in \mathbb{R}_+^K : \bar{\mathbf{p}} \preccurlyeq \mathbf{P}_{max}\right\}$ and $\varpi \in \mathcal{D} \triangleq \left\{\bar{\varpi} \in \mathbb{Z}_{>0} : \bar{\varpi} \le \frac{C_F}{2RB_W}\right\}$, where $\mathbf{P}_{max} \triangleq [P_{1,max},\ldots,P_{K,max}]^T$, the following sufficient and necessary conditions are satisfied:

$$\xi_k^{TL}\left(\mathbf{p},\varpi\right) + \frac{2b_kR\varpi}{C_F\log_2(M)} < \mathcal{T}_k^{th}, \quad \forall k, \tag{28a}$$

$$\sum_{k=1}^{K}\frac{\omega_k}{\mathcal{T}_k^{th} - \xi_k^{TL}\left(\mathbf{p},\varpi\right) - \frac{2B_WR\varpi}{C_F}} \le F_T. \tag{28b}$$

Proof: The individual conditions in (28a) are necessary to ensure that each IoTD can transmit the computation bits to the BBU within the maximum tolerable latency. Subsequently, (28b) guarantees that the total computational resource available at the BBU is enough to assign the computational resource to each IoTD to execute their computational tasks while satisfying the corresponding latency requirement. ■

The conditions in (28) can be enforced by a proper admission control strategy [10], [25], or an appropriate choice of the fronthaul capacity or the BBU computational capability [26]. In what follows, we assume that $\mathcal{P}_3$ is feasible and present the corresponding solution. Accordingly, to solve $\mathcal{P}_3$, we first fix the number of quantization bits ϖ in $\mathcal{P}_3$ and optimize $\mathbf{p}$ and $\mathbf{f}$ by solving the following sub-problem:

$$\mathcal{P}_4: \min_{\{\mathbf{p},\mathbf{f}\}} \mathbf{1}_K^T\mathbf{p} \quad \textit{s.t. } \mathrm{C}_2, \mathrm{C}_3, \quad \mathrm{C}_8: \frac{b_k}{B_W\log_2\left(1+\frac{p_k\alpha_{k,k}}{\eta_k+\sum_{j=1,\ j\neq k}^{K} p_j\alpha_{k,j}+2^{-2\varpi}\sum_{r=1}^{R}\Xi_{r,k}(\mathbf{p})}\right)} + \frac{\omega_k}{f_k} \le \bar{\mathcal{T}}_k^{th}, \quad \forall k, \tag{29}$$

where $\bar{\mathcal{T}}_k^{th} \triangleq \mathcal{T}_k^{th} - \frac{2b_kR\varpi}{C_F\log_2(M)}$. Subsequently, the number of quantized bits ϖ is obtained through the following feasibility problem

$$\mathcal{P}_5: \text{Find } \{\varpi\} \quad \textit{s.t. } \mathrm{C}_4, \mathrm{C}_5, \quad \mathrm{C}_9: \frac{b_k}{B_W\log_2\left(1+\frac{p_k\alpha_{k,k}}{\tilde{\eta}_k+2^{-2\varpi}\sum_{r=1}^{R}\Xi_{r,k}(\mathbf{p})}\right)} + \frac{2b_kR\varpi}{C_F\log_2(M)} \le \tilde{\mathcal{T}}_k^{th}, \quad \forall k, \tag{30}$$

where $\tilde{\mathcal{T}}_k^{th} \triangleq \mathcal{T}_k^{th} - \frac{\omega_k}{f_k}$ and $\tilde{\eta}_k \triangleq \eta_k + \sum_{j=1,\ j\neq k}^{K} p_j\alpha_{k,j}$.

Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on April 29,2022 at 09:05:06 UTC from IEEE Xplore. Restrictions apply.

1) Solution for the Problem $\mathcal{P}_4$: $\mathcal{P}_4$ is still non-convex and difficult to solve due to C_8, which can be expressed as (31), shown at the bottom of this page, such that $g_k(\mathbf{p}, f_k) \triangleq g'_k(\mathbf{p}) + g'_k(f_k)$. C_8 is non-convex due to $g_k(\mathbf{p}, f_k)$. To overcome this difficulty, we exploit the framework of successive inner convexification for $g_k(\mathbf{p}, f_k)$ [44]. The successive inner convexification optimizes a sequence of approximate convex problems, denoted by $\mathcal{A}_{CP}$, which allows the development of a computationally-efficient algorithm converging to a first-order optimal solution [44], [45]. As the non-convexity of $g_k(\mathbf{p}, f_k)$ stems from $g'_k(\mathbf{p})$, in the following we obtain a convex approximation for $g'_k(\mathbf{p})$. To be more specific, letting $p_k = 2^{q_k}$, we have

$$g'_k(\mathbf{q}) = -\log_2\left(1+\frac{2^{q_k}\alpha_{k,k}}{\eta_k+\sum_{j=1,\ j\neq k}^{K} 2^{q_j}\alpha_{k,j}+2^{-2\varpi}\sum_{r=1}^{R}\Xi_{r,k}(\mathbf{q})}\right), \tag{32}$$

where $\mathbf{q} \triangleq [q_1, \ldots, q_K]^T$. In the t-th sequence of convexification, denoted by $\tilde{g}'_k(\mathbf{q}^{(t)}; \mathbf{q}^{(t-1)})$, we require the following three properties to be satisfied for the convex approximation of $g'_k(\mathbf{q}^{(t)})$ [44]:

$$g'_k\left(\mathbf{q}^{(t)}\right) \leq \tilde{g}'_k\left(\mathbf{q}^{(t)}; \mathbf{q}^{(t-1)}\right), \quad \forall t, k, \tag{33a}$$

$$g'_k\left(\mathbf{q}^{(t-1)}\right) = \tilde{g}'_k\left(\mathbf{q}^{(t-1)}; \mathbf{q}^{(t-1)}\right), \quad \forall k, \tag{33b}$$

$$\nabla g'_k\left(\mathbf{q}^{(t-1)}\right) = \nabla \tilde{g}'_k\left(\mathbf{q}^{(t-1)}; \mathbf{q}^{(t-1)}\right), \quad \forall k, \tag{33c}$$

where $\mathbf{q}^{(t-1)}$ is the optimal solution for $\mathcal{A}_{CP}^{(t-1)}$. The central step of this approach is to find a suitable approximation for $g'_k\left(\mathbf{q}^{(t)}\right)$, $\forall k$, which fulfills the requirements in (33), given by the following lemma.

Lemma 2: For a given $\mathbf{q}^{(t-1)} \succeq 0$, a $\tilde{g}'_k\left(\mathbf{q}^{(t)}; \mathbf{q}^{(t-1)}\right)$ that satisfies (33) can be defined as in (34), shown at the bottom of the next page, where

$$\psi_k^{(t-1)} \triangleq \frac{\zeta_k^{(t-1)}}{1+\zeta_k^{(t-1)}},$$

$$\beta_k^{(t-1)} \triangleq \log_2\left(1+\zeta_k^{(t-1)}\right) - \frac{\zeta_k^{(t-1)}}{1+\zeta_k^{(t-1)}}\log_2\left(\zeta_k^{(t-1)}\right),$$

$$\zeta_k^{(t-1)} \triangleq \frac{2^{q_k^{(t-1)}}\alpha_{k,k}}{\bar{\eta}_k\left(\mathbf{q}^{(t-1)}\right)+\sum_{j=1,\ j\neq k}^{K} 2^{q_j^{(t-1)}}\alpha_{k,j}},$$

and

$$\bar{\eta}_k\left(\mathbf{q}^{(t-1)}\right) \triangleq \eta_k + 2^{-2\varpi}\sum_{r=1}^{R}\Xi_{r,k}\left(\mathbf{q}^{(t-1)}\right).$$

Proof: From (32), we have (35), shown at the bottom of the next page, where step (a) is obtained by leveraging the lower-bound of the logarithmic function [45], i.e., $\log_2(1+\zeta) \geq \psi\log_2(\zeta) + \beta$, where $\psi = \frac{\bar{\zeta}}{1+\bar{\zeta}}$ and $\beta = \log_2\left(1+\bar{\zeta}\right) - \frac{\bar{\zeta}}{1+\bar{\zeta}}\log_2\left(\bar{\zeta}\right)$. Hence, $\tilde{g}'_k\left(\mathbf{q}^{(t)}; \mathbf{q}^{(t-1)}\right), \forall k$ satisfies (33a), where (33b) and (33c) hold at $\mathbf{q}^{(t)} = \mathbf{q}^{(t-1)}$. ■

Accordingly from (31), for the t-th sequence, we have

$$\begin{aligned} &g_k\left(\mathbf{q}^{(t)}, f_k^{(t)}\right) \\ &\leq \tilde{g}'_k\left(\mathbf{q}^{(t)}; \mathbf{q}^{(t-1)}\right) + g'_k\left(f_k^{(t)}\right), \\ &= -\psi_k^{(t-1)}\left(\Gamma_k + q_k^{(t)}\right) - \beta_k^{(t-1)} + \frac{f_k^{(t)} b_k}{B_W f_k^{(t)}\bar{\mathcal{T}}_k^{th} - B_W\omega_k} \leq 0, \end{aligned} \tag{36}$$

where

$$\Gamma_k \triangleq \left(\log_2(\alpha_{k,k}) - \log_2\left(\bar{\eta}_k\left(\mathbf{q}^{(t)}\right) + \sum_{j=1,\ j\neq k}^{K}\alpha_{k,j}2^{q_j^{(t)}}\right)\right).$$

As the logarithm of the sum of the exponentials is a convex function [46], $g_k\left(\mathbf{q}^{(t)}, f_k^{(t)}\right)$, $\forall k$ is jointly convex in $\mathbf{q}^{(t)}$ and $f_k^{(t)}$. Consequently, ignoring the sequence index t, the approximate convex problem $\mathcal{A}_{CP}^{(t)}$ for the non-convex problem $\mathcal{P}_4$ is given by

$$\begin{aligned} \mathcal{P}_6: \ &\min_{\{\mathbf{q},\mathbf{f}\}} \ \mathbf{1}_K^T 2^{\circ\mathbf{q}} \\ &\textit{s.t.}\ \mathrm{C}_2,\ \mathrm{C}_{10}: -\psi_k(\Gamma_k + q_k) - \beta_k \\ &\qquad + \frac{f_k b_k}{B_W f_k\bar{\mathcal{T}}_k^{th} - B_W\omega_k} \leq 0, \quad \forall k, \\ &\qquad \mathrm{C}_{11}: 2^{q_k} \leq P_{k,max}, \quad \forall k. \end{aligned} \tag{37}$$

where $2^{\circ\mathbf{q}} \triangleq [2^{q_1}, \ldots, 2^{q_K}]^T$. Further, assuming $\mathbf{q}^{(t)} \preccurlyeq \mathbf{q}^{(t-1)}$[3] and $\exists\ \mathbf{q}^{(t)}$ such that (28) is satisfied, i.e., $2^{q_k^{(t)}} \leq P_{k,max}$, $\forall t, k$, we formulate the following problem

[3]The subsequent derivations in Lemma 3 and Lemma 4 comply with this assumption.

$$\underbrace{\frac{b_k}{B_W\log_2\left(1+\frac{p_k\alpha_{k,k}}{\eta_k+\sum_{j=1,\ j\neq k}^{K}p_j\alpha_{k,j}+2^{-2\varpi}\sum_{r=1}^{R}\Xi_{r,k}(\mathbf{p})}\right)}}_{g_k(\mathbf{p},f_k)} + \frac{\omega_k}{f_k} - \bar{\mathcal{T}}_k^{th} \leq 0,$$

$$\Rightarrow \underbrace{-\log_2\left(1+\frac{p_k\alpha_{k,k}}{\eta_k+\sum_{j=1,\ j\neq k}^{K}p_j\alpha_{k,j}+2^{-2\varpi}\sum_{r=1}^{R}\Xi_{r,k}(\mathbf{p})}\right)}_{g'_k(\mathbf{p})\text{: non-convex}} + \underbrace{\frac{f_k b_k}{B_W f_k\bar{\mathcal{T}}_k^{th} - B_W\omega_k}}_{g'_k(f_k)\text{: convex}} \leq 0 \tag{31}$$

Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on April 29,2022 at 09:05:06 UTC from IEEE Xplore. Restrictions apply.

based on $\mathcal{P}_6$:

$$\mathcal{P}_7: \min_{\{\mathbf{q},\mathbf{f}\}} \mathbf{1}_K^T 2^{\circ\mathbf{q}} \quad s.t.\ \mathrm{C}_2,\ \mathrm{C}_{12}: -\psi_k\left(\bar{\Gamma}_k + q_k\right) - \beta_k + \frac{f_k b_k}{B_W f_k \bar{\mathcal{T}}_k^{th} - B_W \omega_k} \leq 0, \quad \forall k, \tag{38}$$

where

$$\Gamma_k \geq \bar{\Gamma}_k \triangleq \left(\log_2(\alpha_{k,k}) - \log_2\left(\bar{\eta}_k\left(\mathbf{q}^{(t-1)}\right) + \sum_{j=1,\, j\neq k}^{K} \alpha_{k,j} 2^{q_j^{(t-1)}}\right)\right).$$

Therefore, any feasible solution for $\mathcal{P}_7$ is a feasible solution for $\mathcal{P}_6$. Accordingly, in the following we focus on $\mathcal{P}_7$ and resort to the KKT conditions to find the closed-form expressions for $\mathbf{p}^{(t)}$, i.e., $2^{\circ\mathbf{q}^{(t)}}$ and $\mathbf{f}^{(t)}$. Subsequently, the Lagrangian associated with $\mathcal{P}_7$ is given by

$$\Upsilon\left(q_k, f_k, \vartheta_k, \mu\right) = \mathbf{1}_K^T 2^{\circ\mathbf{q}} + \sum_{k=1}^{K} \vartheta_k \left[-\psi_k\left(\bar{\Gamma}_k + q_k\right) - \beta_k + \frac{f_k b_k}{B_W f_k \bar{\mathcal{T}}_k^{th} - B_W \omega_k}\right] + \mu\left(\mathbf{1}_K^T \mathbf{f} - F_T\right), \tag{39}$$

where the variables ϑ_k and μ are the non-negative Lagrange multipliers. Accordingly, the KKT conditions are given by

$$\frac{\Upsilon}{q_k} = (\log 2) 2^{q_k} - \vartheta_k \psi_k = 0, \quad \forall k, \tag{40}$$

$$\frac{\Upsilon}{f_k} = -\frac{B_W \vartheta_k b_k \omega_k}{\left(B_W f_k \bar{\mathcal{T}}_k^{th} - B_W \omega_k\right)^2} + \mu = 0, \quad \forall k, \tag{41}$$

$$\vartheta_k \left[-\psi_k\left(\bar{\Gamma}_k + q_k\right) - \beta_k + \frac{f_k b_k}{B_W f_k \bar{\mathcal{T}}_k^{th} - B_W \omega_k}\right] = 0, \quad \vartheta_k \geq 0, \forall k, \tag{42}$$

$$\mu\left(\mathbf{1}_K^T \mathbf{f} - F_T\right) = 0, \quad \mu \geq 0. \tag{43}$$

Since $\mathcal{P}_7$ satisfies (28), all the K IoTDs are served, i.e., $q_k > 0$ and $f_k > 0,\ \forall k$. Accordingly, the conditions (40), (41) and (43) imply $\vartheta_k > 0,\ \forall k$ and $\mu > 0$, which means that the computational capability at the BBU server is fully utilized, i.e.,

$$\mathbf{1}_K^T \mathbf{f} = F_T. \tag{44}$$

From another perspective, we can also obtain that $\mathbf{1}_K^T \mathbf{f} < F_T$ would be sub-optimal, since at least the value of one p_k can be further reduced by increasing the value of the corresponding f_k. Furthermore, $\vartheta_k > 0,\ \forall k$ implies that the latency constraint is always active, i.e.,

$$-\psi_k\left(\bar{\Gamma}_k + q_k\right) - \beta_k + \frac{f_k b_k}{B_W f_k \bar{\mathcal{T}}_k^{th} - B_W \omega_k} = 0. \tag{45}$$

This equation establishes a one-to-one relationship between the transmit power $p_k = 2^{q_k}$ and the number of cycles per second f_k at the BBU server assigned to the k-th IoTD. Consequently, from (40) and (41), we obtain the expression for the optimal computational resource f_k as

$$f_k = \frac{1}{\bar{\mathcal{T}}_k^{th}}\left[\sqrt{\frac{(\log 2) b_k \omega_k 2^{q_k}}{\mu B_W \psi_k}} + \omega_k\right]. \tag{46}$$

By substituting (46) into (44) to obtain μ and by replacing 2^{q_k} with p_k, f_k is further transformed into (47), shown at the bottom of the next page. Finally, from (45), the optimal transmit power for the k-th IoTD is given by

$$p_k = 2^{\left[\frac{1}{\psi_k}\left(\frac{f_k b_k}{B_W \bar{\mathcal{T}}_k^{th} f_k - B_W \omega_k} - \beta_k\right) - \bar{\Gamma}_k\right]}. \tag{48}$$

Lemma 3: Under the assumption that $\exists\, \mathbf{p}$ such that (28) is satisfied, $p_k, \forall k$ obtained by (48) will converge to an optimal solution to $\mathcal{P}_6$ for a given $f_k,\ \forall k$.

Proof: Refer to Appendix.

Lemma 4: Under the assumption that $\exists\, \mathbf{p}$ such that (28) is satisfied, $f_k, \forall k$ given by (47) converges to a KKT point of $\mathcal{P}_6$.

Proof: According to ***Lemma 3***, when $\exists\ \mathbf{p}$ such that (28) is satisfied, $\mathbf{p}$ obtained by (48) converges, i.e., $\mathbf{p}^{(t)} = \mathbf{p}^{(t-1)}$. Accordingly, upon convergence equality holds for $\bar{\Gamma}_k \leq \Gamma_k,\ \forall k$, which results in the equivalent KKT conditions for $\mathcal{P}_6$ and $\mathcal{P}_7$. Hence, $f_k,\ \forall k$ given by (47) converge to KKT point of $\mathcal{P}_6$. ■

For clarity, we summarize the above procedure in Algorithm 1, which describes the framework to obtain the transmit power and computational resource for the K IoTDs. Since $\mathcal{P}_6$ satisfies the conditions in (33), its solution will converge to the KKT point of $\mathcal{P}_4$, which accordingly gives a local minimum of $\mathcal{P}_4$ [44, Corollary 1]. Hence, according to ***Lemma 3*** and ***Lemma 4***, Algorithm 1 converges to a local minimum of $\mathcal{P}_4$.

2) Solution for the Problem $\mathcal{P}_5$: $\mathcal{P}_5$ is a non-convex problem due to C_5 and C_9. Noting that there is only a single integer variable to be optimized, we resort to the line-search method to find the optimal ϖ over the feasible set. Accordingly, ϖ is given by (49), shown at the bottom of the next page.

$$\tilde{g}_k'\left(\mathbf{q}^{(t)}; \mathbf{q}^{(t-1)}\right) \triangleq -\psi_k^{(t-1)}\left(\log_2(\alpha_{k,k}) + q_k^{(t)} - \log_2\left(\bar{\eta}_k\left(\mathbf{q}^{(t)}\right) + \sum_{j=1,\, j\neq k}^{K} \alpha_{k,j} 2^{q_j^{(t)}}\right)\right) - \beta_k^{(t-1)} \tag{34}$$

$$g_k'(\mathbf{q}^{(t)}) \overset{(a)}{\leq} -\psi_k^{(t-1)}\left(\log_2(\alpha_{k,k}) + q_k^{(t)} - \log_2\left(\bar{\eta}_k\left(\mathbf{q}^{(t)}\right) + \sum_{j=1,\, j\neq k}^{K} \alpha_{k,j} 2^{q_j^{(t)}}\right)\right) - \beta_k^{(t-1)}, = \tilde{g}_k'\left(\mathbf{q}^{(t)}; \mathbf{q}^{(t-1)}\right), \quad \forall k \tag{35}$$

Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on April 29,2022 at 09:05:06 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1 Iterative Algorithm to Solve $\mathcal{P}_6$

1: **Input**: $\mathbf{p}^{(0)}$, $\varpi^{(0)}$.
2: Initialize $t \leftarrow 1$;
3: **repeat**
4: Update $f_k^{(t)}$ using (47), $\forall k$;
5: Update $p_k^{(t)}$ using (48), $\forall k$;
6: $t \leftarrow t+1$.
7: **until** convergence
8: **Output**: p_k, f_k, $\forall k$.

Algorithm 2 Overall Algorithm to Solve $\mathcal{P}_2$

1: **Input**: $\mathbf{p}^{(0)}$, $\varpi^{(0)}$.
2: Initialize $t \leftarrow 1$;
3: **repeat**
4: Update $\mathbf{W}^{(t)}$ using (23);
5: Update $\mathbf{p}^{(t)}$ and $\mathbf{f}^{(t)}$ using Algorithm 1;
6: Update $\varpi^{(t)}$ using (49);
7: $t \leftarrow t+1$.
8: **until** convergence
9: **Output**: $\mathbf{W}$, $\mathbf{p}$, $\mathbf{f}$, ϖ.

3) Overall Algorithm for the Problem $\mathcal{P}_2$: Algorithm 2 summarizes the overall algorithm to solve $\mathcal{P}_2$. Specifically, for a given feasible $\mathbf{p}$, $\mathbf{f}$ and ϖ, the algorithm starts by obtaining $\mathbf{W}$ using (23). Subsequently, for the obtained $\mathbf{W}$ and a fixed ϖ, $\mathbf{p}$ and $\mathbf{f}$ are updated using Algorithm 1. Finally, for the obtained $\mathbf{W}, \mathbf{p}$, and $\mathbf{f}$, we find a feasible ϖ using (49) for the next iteration. As the objective of $\mathcal{P}_2$ is decreasing in each iteration owing to Algorithm 1, Algorithm 2 converges to a local minimum.

IV. Low-Complexity Implementation for the Joint Optimization Based on DL

Although Algorithm 2 obtains near-optimal solutions for $\mathcal{P}_2$, it involves an interleaved loop structure that can limit its practicability in terms of real-time processing. Accordingly, in this section, we present a supervised DL method using a DNN to obtain the solution of $\mathcal{P}_2$, more efficiently by passing the input parameters of $\mathcal{P}_2$ through the DNN. The DNN is trained offline and then used as an approximation of the proposed optimization algorithm [35]. To simplify the proposed DNN framework, we note that for an optimal $\{\mathbf{p}, \varpi\}$ obtained through solving $\mathcal{P}_3$, $\mathbf{W}$ can be computed directly using the closed-form solution given in (23). Accordingly, based on the universal approximation theorem [34], which states that any deterministic continuous function (or optimal policy) defined over a compact set can be approximated arbitrarily well with a DNN, we deploy a DNN to approximate the optimal policy of $\mathcal{P}_3$, denoted by Θ:

$$\Theta : \mathbf{x} \rightarrow \mathbf{y}, \tag{50}$$

which maps the input features $\mathbf{x}$ to the optimal resource allocation $\mathbf{y}$ [32], [33], [38]. To be specific, the computation bits, the latency thresholds and the effective channel gains for each IoTD, given by $\mathbf{b} \triangleq [b_1, \ldots, b_k]^T, \forall k$, $\mathbf{T}_{th} \triangleq \left[\mathcal{T}_1^{th}, \ldots, \mathcal{T}_K^{th}\right]^T, \forall k$ and $\mathbf{c} \triangleq [c_1, \cdots, c_K]^T = |(\mathbf{VH} \odot \mathbf{I})\, \mathbf{1}_K|$, respectively, and jointly denoted by $\mathbf{x} \triangleq \left[\mathbf{b}^T, \mathbf{T}_{th}^T, \mathbf{c}^T\right]^T$, are mapped to $\mathbf{y} \triangleq \left[\mathbf{p}^T, \mathbf{f}^T, \varpi\right]^T$.

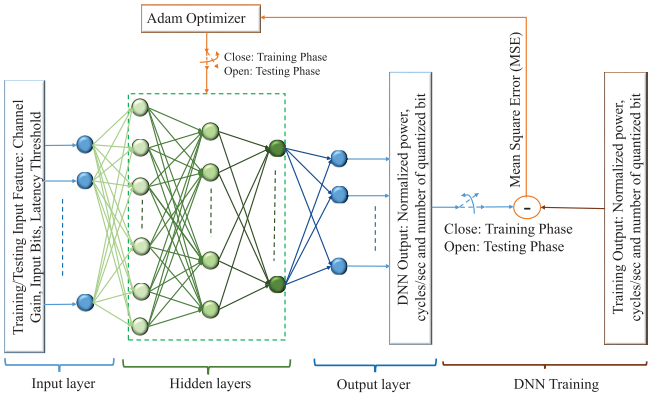


Fig. 2. DNN architecture for the proposed supervised DL with the training and testing phase.

A. DNN Framework

Next, we describe the DNN architecture used in our work, as shown in Fig. 2. Specifically, the DNN consists of a) one input layer of $3K$ neurons given by $\mathbf{x}$, b) one output layer of $2K+1$ neurons corresponding to $\mathbf{y} \triangleq \left[\mathbf{p}^T, \mathbf{f}^T, \varpi\right]^T$, and c) $L-1$ fully connected hidden layers with $l=0$ and $l=L$ denoting the input and output layers, respectively. The number of neurons in each layer is denoted by n_l, and accordingly, we have $n_0 = 3K$ and $n_L = 2K+1$. For each hidden layer l, the output $\mathbf{o}_l \in \mathbb{R}^{n_l \times 1}$ is calculated as

$$\mathbf{o}_l = \mathrm{ReLU}\left[\mathbf{Q}_l \left(\mathbf{o}_{l-1} \odot \mathbf{r}_l\right) + \mathbf{b}_l\right], \quad l \in \{1, \ldots, L\}, \tag{51}$$

where $\mathbf{o}_{l-1} \in \mathbb{R}^{n_{l-1} \times 1}$ is the output of the $(l-1)$-th layer with $\mathbf{o}_0 = \mathbf{x}$, $\mathbf{Q}_l \in \mathbb{R}^{n_l \times n_{l-1}}$ and $\mathbf{b}_l \in \mathbb{R}^{n_l \times 1}$ are respectively the

$$f_k = \frac{1}{\tilde{\mathcal{T}}_k^{th}} \left[\frac{F_T - \sum_{k=1}^{K} \frac{\omega_k}{\tilde{\mathcal{T}}_k^{th}}}{\sum_{k=1}^{K} \frac{1}{\tilde{\mathcal{T}}_k^{th}} \sqrt{\frac{(\log 2) b_k \omega_k p_k}{B_W \psi_k}}} \sqrt{\frac{(\log 2) b_k \omega_k p_k}{B_W \psi_k}} + \omega_k \right] \tag{47}$$

$$\varpi = \underset{\tilde{\varpi}}{\mathrm{argmax}} \left\{ \tilde{\varpi} \in \mathbb{Z}_{>0} : \frac{b_k}{B_W \log_2\left(1 + \frac{p_k \alpha_{k,k}}{\tilde{\eta}_k + 2^{-2\tilde{\varpi}} \sum_{r=1}^{R} \Xi_{r,k}(\mathbf{p})}\right)} + \frac{2 b_k R \varpi}{C_F \log_2(M)} \leq \tilde{\mathcal{T}}_k^{th}, \forall k, \tilde{\varpi} \leq \frac{C_F}{2 B_W R} \right\} \tag{49}$$

Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on April 29,2022 at 09:05:06 UTC from IEEE Xplore. Restrictions apply.

weight matrix and bias vector at the l-th layer, $\mathbf{r}_l \in \mathbb{Z}^{n_{l-1}\times 1}$ is a vector of independent Bernoulli random variables each of which has a probability $(1-\nu)$ of being 1, ν denotes the dropout rate [47], and $\mathrm{ReLU}(.) = \max(.,0)$ is the Rectified Linear Unit function, which introduces nonlinearity to the network. Note that the dropout regularization enforced by $\mathbf{r}_l,\ \forall l = 1,\cdots,L-1$, prevents the overfitting of the DNN to the training data samples and improves its generalization performance [47]. Accordingly, the DL method involves

1) Obtaining the training data samples, i.e., the training input and corresponding output, denoted by $\{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}, i = 1,\cdots Z$, where Z is the number of training data samples, by running Algorithm 2 offline.
2) Normalizing the training data samples such that $\{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\} \in [0,1]\ \forall i$.
3) Deploying the mini-batch gradient descent based on Adam optimizer to train the DNN [33], [33], [40], [48], [49] as shown in Fig. 2 (Training phase), which effectively minimizes the mean square error (MSE), defined by $\mathrm{MSE} \triangleq \frac{\sum_{b=1}^{B_M}\sum_{i=1}^{2K+1}(o_{L,i,b}-y_{i,b})^2}{B_M(2K+1)}$, where B_M is the number of mini-batches, $o_{L,i,b}$ and $y_{i,b}$ are the outputs at the i-th neuron of the L-th layer and the corresponding training output, respectively, for the b-th mini-batch.
4) After the training phase, the DNN is used to obtain the desired output $\{\mathbf{p}, \mathbf{f}, \varpi\}$ based on the test input data, i.e., $\{\mathbf{b}, \mathbf{T}_{th}, \mathbf{c}\}$, which can be real-world data from the C-RAN network, as shown in Fig. 2 (Testing phase). Note that dropout regularization is removed during the testing phase [47]. Furthermore, with $\mathbf{V}$ obtained from (20) and (21) and $\{\mathbf{p}, \varpi\}$ obtained from the trained DNN, $\mathbf{W}$ is obtained using (23).

B. Enhancing the Adaptability of the DNN With DTL

As we will demonstrate in the next section, the proposed DNN framework provides a close approximation for the optimal policy of $\mathcal{P}_3$, however, the practical implementation of the DNN framework is still limited due to the following reason. There are hidden parameters that are not included in the input of the DNN but impact the optimal solutions, such as $\{B_W, F_T, C_F\}$. Furthermore, these variables are inherently dynamic, depending on the operating scenario. Accordingly, the pre-trained DNN can suffer from performance deterioration as the operating parameters change. This issue is identified as task mismatch, i.e., the test scenario is different from the trained one [37], [38], [50]. A natural way to resolve this issue is to train the DNN from scratch for the updated scenario after collecting enough additional training data samples. Nevertheless, it is practically infeasible to train the DNN from scratch, given the requirement of large training data samples coupled with the corresponding long training time.

In this case, DTL can be leveraged by transferring knowledge from a related operating scenario with the availability of enough training data samples, denoted by $\mathbf{x}^{(A,i)},\ i = 1,\ldots,Z_A$, where Z_A is the number of training data samples. The idea is that although the updated scenario is different from the initial scenario, they share the same structure of the underlying optimization problem. Specifically, the transfer of the knowledge is implemented by first training a DNN under the initial scenario with $\mathbf{x}^{(A,i)},\ \forall i$, yielding a tentative configuration of the DNN parameters. Next, for an updated scenario, a second training phase is performed with limited training data samples, denoted by $\mathbf{x}^{(B,i)},\ i = 1,\ldots,Z_B$, where $Z_B \ll Z_A$, by *reusing* and *refining* the configuration of the weights and bias from the initial training phase as the initialization point for the training algorithm as shown in Fig. 3. Note that while *reusing* refers to fixing the parameters in the first few layers, which captures the common features between the initial and updated scenarios, *refining* denotes fine-tuning the parameters in the last few-layers with $\mathbf{x}^{(B,i)},\ \forall i$ to learn the optimal policy specific to the updated scenario [37], [50]. Consequently, DTL achieves a good performance under the updated scenario, thereby enhancing the adaptability and the practicability of the DNN.

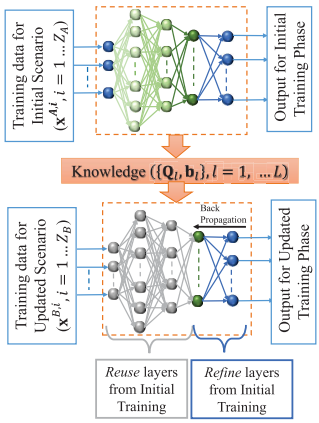


Fig. 3. DTL via reusing and refining the parameters of a pre-trained DNN.

V. Numerical Results

In this section, we evaluate the performance of our proposed approach via Monte-Carlo simulations. Unless otherwise stated, we consider a network composed of $N = 128$ antennas randomly deployed on a wall in a $10\ \mathrm{m}\times 10\ \mathrm{m}\times 10\ \mathrm{m}$ indoor room as shown in Fig. 4. Furthermore, there are $K = 10$ single-antenna IoTDs uniformly distributed inside the room. The number of bits b_k and the latency threshold $\mathcal{T}_k^{th}$ for each IoTD's computational task C_k are randomly assigned between 10 kbs to 20 kbs and 0.5 s to 1 s, respectively. The computation bits are encoded using the QPSK modulation, i.e., $M = 4$. For the sake of simplicity, the number of CPU cycles needed for completing C_k is set as a linear function of b_k, i.e., $\omega_k = \eta b_k$, with $\eta = 50$ [10]. The carrier frequency of the wireless links is taken to be $f_\varepsilon = \frac{\varepsilon}{\lambda} = 1.5$ GHz with a transmission bandwidth of $B_W = 180$ KHz, where $\varepsilon = 3\times 10^8$ m/s. Furthermore, the channel parameters are given as $\delta = 3.7$, $\kappa = (13 - 0.03\, d_{n,k}[m])\,\mathrm{dB}$ and $\sigma_{shad} = 6\ \mathrm{dB}, \forall n,k$ [8], [32], [51]. The transmit power constraint for each IoTD is $P_{k,max} = 0$ dBm. The power spectral density of the background noise at the xL-MIMO RRH is assumed to be -169 dBM/Hz, and the noise figure due to the receiver processing is 7 dB [9]. Lastly, it is

Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on April 29,2022 at 09:05:06 UTC from IEEE Xplore. Restrictions apply.

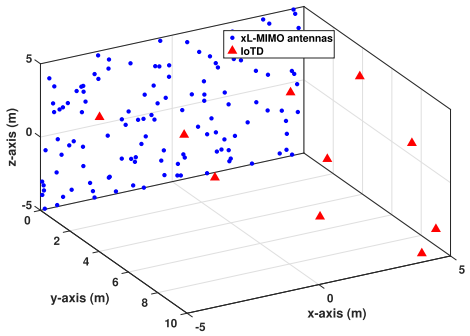


Fig. 4. Simulation set-up with N antennas (blue circles) deployed on a wall and K IoTDs (red triangles) distributed in an indoor room.

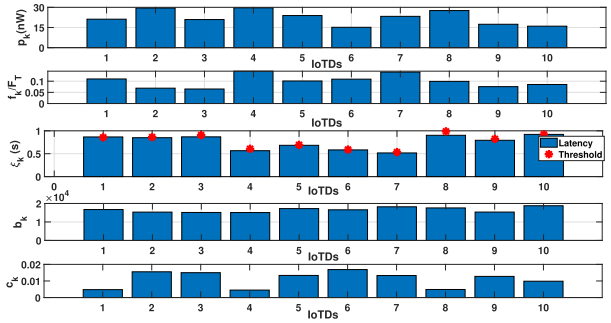


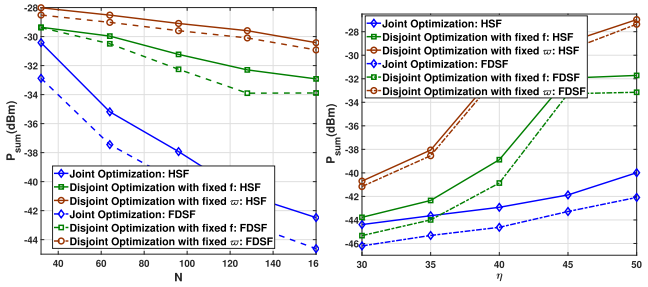
Fig. 5. Optimal transmitted power p_k, normalized CPU cycles f_k/F_T and overall latency ξ_k, w.r.t. the latency threshold $\mathcal{T}_k^{th}$, the number of transmit bits b_k and the effective channel gain c_k corresponding to each IoTD.

assumed that the BBU server has a computational capability of $F_T = 15$ MHz cycles/s with a fronthaul capacity of $C_F = 100$ MHz. The above choice of parameters guarantees the non-emptiness of the feasible set for $\mathcal{P}_3$, where $p_k^{(0)}, \forall k$ and $\varpi^{(0)}$ are selected randomly within the feasible sets Ψ and $\mathcal{D}$, respectively.

A. Performance per IoTDs' Locations and Channel Realizations

To gain insights from the communication and computational resource allocations by the proposed algorithm, we firstly consider the resource allocation for a set of particular IoTDs' locations and channel realizations. In Fig. 5, we illustrate the obtained communication (transmit power $p_k, \forall k$, first sub-figure) and computational (normalized number of CPU cycles $f_k/F_T, \forall k$, second sub-figure) resources assigned to each IoTD with respect to (w.r.t.) the corresponding latency thresholds ($\mathcal{T}_k^{th}, \forall k$, third sub-figure), the number of computation bits ($b_k, \forall k$, fourth sub-figure) and effective channel gains at the BBU ($c_k, \forall k$, fifth sub-figure). In the third sub-figure, we also plot the overall latency $\xi_k, \forall k$ computed using (16).

As observed, the proposed algorithm assigns a higher transmit power and CPU cycles to IoTDs with a poor effective channel gain (IoTD 8), a larger number of computation bits (IoTD 2) or a stringent latency constraint (IoTD 7). An interesting observation is that, with similar channel gains, the latency constraint dominates over the number of computation bits in determining the allocation of the communication and



(a) P_{sum} vs N (b) P_{sum} vs η

Fig. 6. Total transmit power P_{sum} versus the number of antennas N and the computational load η for the proposed joint optimization and the disjoint optimizations.

computational resources as observed for IoTD 1 and 4. This demonstrates that the latency constraints play a crucial role in the computation-offloading for the IoTDs. Furthermore, it is seen that the computational tasks of all the IoTDs are executed within the respective latency constraint.

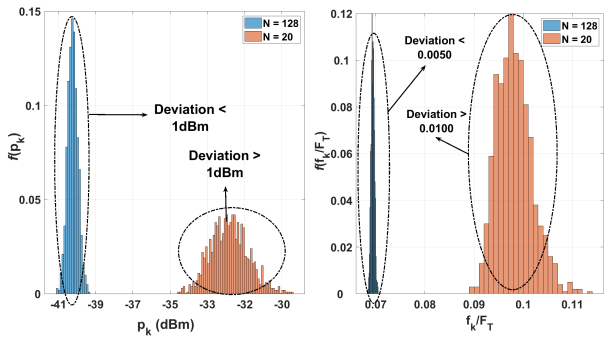
B. Joint Versus Disjoint Optimization

In this section, we evaluate the merit of the proposed algorithm with two benchmark algorithms: 1) Disjoint optimization with fixed $\mathbf{f}$: Solving $\mathcal{P}_2$ with Algorithm 2 where $p_k, \forall k$ and ϖ are optimized with $f_k = \frac{\omega_k F_T}{\sum_{k=1}^{K} \omega_k}, \forall k$, which meets the computational rate constraint F_T with equality [11], and 2) Disjoint optimization with fixed ϖ: Solving $\mathcal{P}_2$ with Algorithm 2 where $p_k, \forall k$ and $f_k, \forall k$ are optimized with the number of quantization bits ϖ fixed at $\varpi = \left\lfloor \frac{C_F}{4B_W L} \right\rfloor$, i.e., half of the maximum feasible ϖ. We assess the usefulness of the algorithms w.r.t. the number of antennas N at the xL-MIMO RRH and the computational load given by the ratio $\eta = \frac{\omega_k}{b_k}$ between the required number of CPU cycles ω_k and number of computation bits b_k [11].

Fig. 6(a) shows the total transmit power of the IoTDs w.r.t. N for $\eta = 50$, obtained using Algorithm 2 and the disjoint optimization algorithms, with both the HSF and the FDSF. It can be observed that the proposed joint optimization algorithm yields a considerable gain compared to the disjoint optimization algorithms, where deploying a large number of antennas results in a decrease in the total transmit power. This decrease in the total transmit power is because of the channel gain, which is proportional to N, resulting in a decrease in the required transmit power of each IoTD [6]. Furthermore, this explains the use of the xL-MIMO with a large N to minimize the power drainage of IoTDs and consequently, extend their battery life.

Next, Fig. 6(b) presents the total transmit power of the IoTDs w.r.t. η for $N = 128$ and $\omega_k = \eta b_k, \ \forall k$, obtained using the algorithms, with both the HSF and the FDSF. Specifically, η is varied with b_k and $\mathcal{T}_k^{th}$ randomly set between 10 kbs to 20 kbs and 0.5 s to 1 s, respectively. It can be observed that the proposed joint optimization algorithm outperforms the disjoint optimization algorithms for the computational tasks with a stringent computational requirement. Finally, it can be seen from Fig. 6 that there is a performance loss for the HSF

Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on April 29,2022 at 09:05:06 UTC from IEEE Xplore. Restrictions apply.



(a) pdf for p_k (b) pdf for f_k/F_T

Fig. 7. Probability density function of the optimal transmit power p_k and the normalized CPU cycles f_k/F_T assigned to an IoTD.

compared to the FDSF owing to a loss in the spectral efficiency for the hybrid architecture [12], [21]–[24].

C. Deep Neural Network Evaluation

1) Impact of Small-Scale Fading on DNN Training: We begin by evaluating the impact of the small-scale fading on the resource allocation for the computation-offloading. Accordingly, for a fixed number of computation bits $(b_k, \forall k)$ and latency thresholds $(\mathcal{T}_k^{th}, \forall k)$, in Fig. 7(a) and 7(b), we show the pdf for the optimal transmit power p_k and normalized number of CPU cycles f_k/F_T assigned to an IoTD across 10^3 channel realizations. It can be seen that for $N = 128$, ${p_k}^4$ and f_k/F_T have a significantly lesser deviation compared to that for $N = 20$. Hence, for a fixed b_k and $\mathcal{T}_k^{th}, \forall k$, these results demonstrate that the proposed HSF with a large number of antennas at the xL-MIMO RRH reduces the impact of the small-scale fading on the resource allocation for the IoTDs as explained in Section 3. Consequently, the BBU needs to update the operating parameters depending only on the large-scale fading of the IoTDs. Additionally, this explains the use of the effective channel at the BBU, representing the large-scale fading corresponding to each IoTD as shown in (19), as an input parameter to train the proposed DNN along with b_k and $\mathcal{T}_k^{th}, \forall k$.

2) DNN Training and Testing: We implemented the proposed DNN scheme with the Keras machine learning toolkit [33], [48], [49]. Accordingly, we consider three hidden layers with 256, 512 and 256 neurons for $l = 1, 2$ and 3, respectively with $\nu = 0.1$. We collected 20000 training data sets, which are split in the ratio of $9 : 1$ for the training and testing of the DNN. Fig. 8(a) shows the training and testing losses w.r.t. the number of epochs, which can be seen to converge within 30 epoch. Accordingly, in Fig. 8(b), we show the communication (transmit power p_k, $\forall k$, first sub-figure) and computational (normalized number of CPU cycles f_k/F_T, $\forall k$, second sub-figure) resources, and the overall latency along with the respective latency thresholds (ξ_k and $\mathcal{T}_k^{th}$, $\forall k$, third sub-figure) obtained from the proposed joint optimization and the trained DNN for two sets of different IoTDs' locations and channel realizations. As observed, the DNN is able to approximately emulate the optimal policy of $\mathcal{P}_3$ to obtain $\{\mathbf{p}, \mathbf{f}, \varpi\}$ while satisfying the latency requirement for each IoTD.

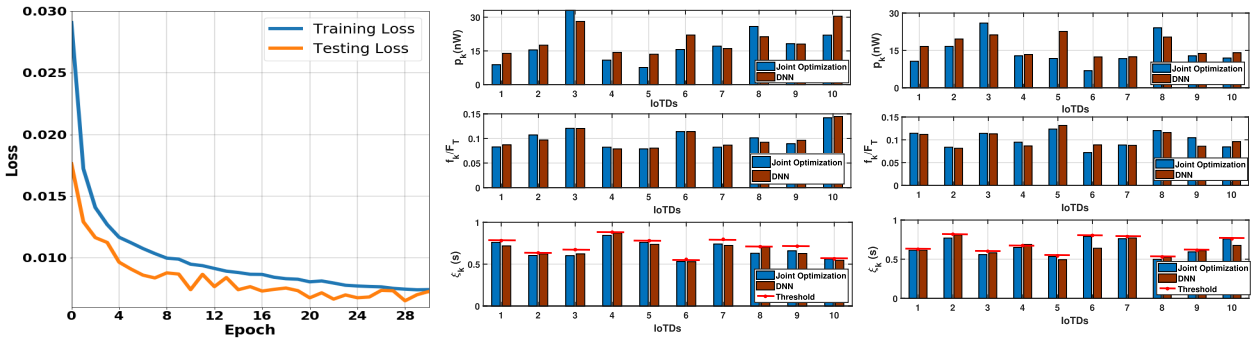
Finally, we evaluate the performance of the DNN method w.r.t. the disjoint optimization algorithms described in the previous sub-section. Accordingly, in Fig. 9(a), we plot the optimal power and computational resource allocations for each IoTDs along with their corresponding latency performance obtained with the proposed joint optimization, the trained DNN and the disjoint optimizations for a set of particular IoTDs' locations and channel realizations, where all the operating parameters are kept constant for the different algorithms. It can be observed that the transmit power of each IoTD obtained with the trained DNN are closer to that obtained from the proposed joint optimization, as compared to the disjoint optimizations. To further consolidate the above observation, in Fig. 9(a), we show the cumulative distribution function (CDF) for P_{sum} obtained with 5000 new testing data sets [35]. It can be observed that P_{sum} of the IoTDs obtained from the trained DNN is close to that obtained from the proposed joint optimization, while significantly outperforming the disjoint optimizations. Moreover, we measured the elapsed time for the computation of the optimal resource allocations through the proposed joint optimization and the trained DNN, where Intel core i7-6700 CPU@3.40 GHz and 16.00 GB RAM are used. The average elapsed time per computation corresponding to the proposed joint optimization and the trained DNN was found to be 118 ms and 1 ms, respectively, which highlights the practicability of the proposed DL method.

3) Adaptability of the DNN With DTL: Based on DTL, we show that the trained DNN can be fine-tuned to approximate the optimal policy of $\mathcal{P}_3$ when the operating parameters are updated. Specifically, the values for $\{B_W, F_T, C_F\}$ are updated to $\{150\text{ KHz}, 20\text{ MHz cycles/s}, 90\text{ MHz}\}$ and the number of IoTDs is changed to $\tilde{K} = 5$. With DTL, the first two hidden layers of the pre-trained DNN are fixed, while the third hidden layer is refined for the updated scenario. Furthermore, as K is changed to $\tilde{K} = K/2$ in the updated scenario, the dimension of the input and output layers changes. Consequently, the unused inputs are replaced by $K/2$ zeros [35], while replacing the output layer consisting of $n_L = 2K + 1$ neurons with a layer of $n_L = 2\tilde{K} + 1$ neurons. Note that the output layer does not need to be replaced if K remains constant between the initial and updated scenarios.

Similar to Fig. 8(a), Fig. 10(a) shows the training and testing losses w.r.t. the number of epochs for two different training procedure: 1) training from scratch with random initialization (RI), and 2) training with DTL, while using only 1000 training data samples. It can be observed that with DTL, the DNN training achieves a significant lower training and testing losses compared to when the DNN is trained with RI. Finally, inline with Fig. 8(b), in Fig. 10(b), we plot the resource allocation and the latency performance for $\tilde{K} = 5$ IoTDs for two sets of different IoTDs' locations and channel realizations. It is observed that even with 1000 training data samples, the DNN with DTL provides a close approximation of the optimal resource allocation, while satisfying the latency requirements.

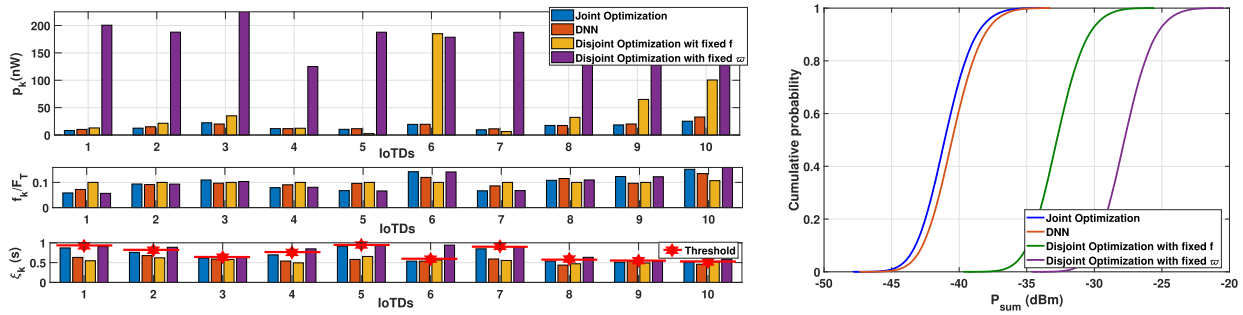
[4]Note that a larger N decreases the required transmit power of the IoTD, thereby further minimizing the total transmit power [6].

Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on April 29,2022 at 09:05:06 UTC from IEEE Xplore. Restrictions apply.



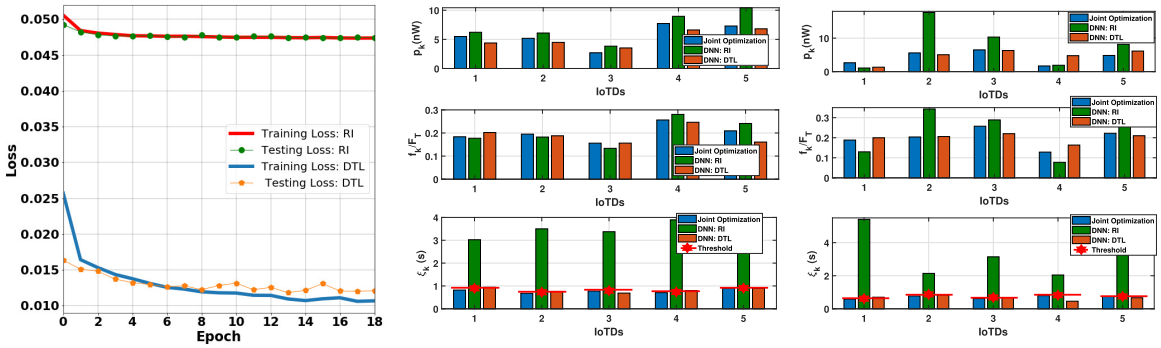
(a) DNN training (b) Resource allocation for IoTDs

Fig. 8. a) Training and testing losses versus epoch for the DNN based learning, and b) $\{p_k, f_k/F_T, \xi_k\}, \forall k$ obtained from the proposed algorithm and the DNN for two sets of different IoTDs' locations and channel realizations.



(a) $\{p_k, f_k/F_T, \xi_k\}, \forall k$ (b) CDF for P_{sum}

Fig. 9. $\{p_k, f_k/F_T, \xi_k\}, \forall k$ obtained from the proposed algorithm, the DNN and the disjoint optimizations for a set of particular IoTDs' locations and channel realizations, and b) CDF for P_{sum} achieved by the proposed joint optimization (average elapsed time: 118ms), the DNN (average elapsed time: 1ms) and the disjoint optimizations for 5000 testing data samples.



(a) DNN training based on DTL (b) Resource allocation with and without DTL

Fig. 10. a) Training and testing losses versus epoch for the DNN with DTL, and b) $\{p_k, f_k/F_T, \xi_k\}, \forall k$ obtained from the proposed algorithm and the DNN with and without DTL for two sets of different IoTDs' locations and channel realizations.

On the other hand, the DNN with RI performs worse than that with DTL in terms of resource utilization efficiency and latency. This result suggests a pre-trained DNN requires very few training data samples for adapting to an updated operating scenario with DTL, which makes the proposed approach very attractive for practical use.

VI. Conclusion

In this paper, we have formulated a computation-offloading problem for IoT applications with a latency constraint in an uplink xL-MIMO C-RAN. The constructed optimization problem that minimizes the total transmit power of the IoTDs while satisfying the latency requirement is found to be non-convex. With the HSF matrix obtained locally at the xL-MIMO RRH, the joint optimization on the baseband combiner, the communication and computational resource allocations, and the number of quantization bits at the BBU is solved with the alternating optimization based on the concepts of the MMSE metric, the successive inner convexification and the linear-search method, respectively. Furthermore, a supervised DL method using the DNN is deployed as an efficient low-complexity solution, while the DTL is adopted to enhance

Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on April 29,2022 at 09:05:06 UTC from IEEE Xplore. Restrictions apply.

the adaptability of the DNN. Numerical results validate the effectiveness of the proposed joint optimization scheme, which outperforms the two benchmarks based on disjoint optimization. The efficiency of the DNN with DTL is also verified.

APPENDIX

For a given $f_k, \forall k$, $\mathcal{P}_6$ can be further transformed into,

$$\mathcal{P}_8: \min_{\{\mathbf{q} \preccurlyeq \log_2(\mathbf{P}_{max})\}} \mathbf{1}_K^T 2^{\circ \mathbf{q}}, \quad s.t.\ \mathrm{C}_{10}. \tag{52}$$

By substituting $p_k = 2^{q_k}$ and assuming $\exists\ \mathbf{p}$ such that (28) is satisfied, i.e., $\mathbf{p} \preccurlyeq \mathbf{P}_{max}$, $\mathcal{P}_8$ reduces to the following problem:

$$\mathcal{P}_9: \min_{\{\mathbf{p} \succeq \mathbf{I}(\mathbf{p})\}} \mathbf{1}_K^T \mathbf{p}, \tag{53}$$

where the function $\mathbf{I}(\mathbf{p}) \triangleq [I_1(\mathbf{p}), \ldots, I_K(\mathbf{p})]^T \in \mathbb{R}^{K \times 1}$ is a *standard interference function* [52], with each entry given by

$$I_k(\mathbf{p}) = 2^{\left[\frac{1}{\psi_k}\left(\frac{f_k b_k}{B_W \bar{\mathcal{T}}_k^{th} f_k - B_W \omega_k} - \beta_k\right) - \Gamma_k\right]}. \tag{54}$$

$\mathcal{P}_9$ is the well-known power control problem [52], which has an optimal solution obtained through the *standard power control algorithm*, given by $\mathbf{p}^{(t)} = \mathbf{I}\left(\mathbf{p}^{(t-1)}\right)$. According to [9, Corollary 4.1 and 4.2], given a feasible $\mathcal{P}_8$, $\mathbf{p}^{(t)} = \mathbf{I}\left(\mathbf{p}^{(t-1)}\right) \preccurlyeq \mathbf{p}^{(t-1)}$ will converge to the optimal solution to $\mathcal{P}_8$ with any initial point $\mathbf{p}^{(0)} \succeq 0$. Hence, (48) which is in the form of $\mathbf{p}^{(t)} = \mathbf{I}\left(\mathbf{p}^{(t-1)}\right)$ converges to an optimal solution for $\mathcal{P}_6$ for a given $f_k, \forall k$. ■

REFERENCES

[1] S. Popli, R. K. Jha, and S. Jain, "A survey on energy efficient narrowband Internet of Things (NBIoT): Architecture, application and challenges," *IEEE Access*, vol. 7, pp. 16739–16776, 2019.

[2] L. Feltrin *et al.*, "Narrowband IoT: A survey on downlink and uplink perspectives," *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 78–86, Feb. 2019.

[3] C. She, C. Yang, and T. Q. S. Quek, "Radio resource management for ultra-reliable and low-latency communications," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 72–78, Jun. 2017.

[4] K. Kumar, J. Liu, Y.-H. Lu, and B. Bhargava, "A survey of computation offloading for mobile systems," *Mobile Netw. Appl.*, vol. 18, no. 1, pp. 129–140, Feb. 2013, doi: 10.1007/s11036-012-0368-0.

[5] C. She, Y. Duan, G. Zhao, T. Q. S. Quek, Y. Li, and B. Vucetic, "Cross-layer design for mission-critical IoT in mobile edge computing systems," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 9360–9374, Dec. 2019.

[6] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.

[7] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.

[8] M. Jung, W. Saad, Y. Jang, G. Kong, and S. Choi, "Performance analysis of large intelligent surfaces (LISs): Asymptotic data rate and channel hardening effects," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2052–2065, Mar. 2020.

[9] L. Liu and R. Zhang, "Optimized uplink transmission in multi-antenna C-RAN with spatial compression and forward," *IEEE Trans. Signal Process.*, vol. 63, no. 19, pp. 5083–5095, Oct. 2015.

[10] S. Barbarossa, S. Sardellitti, and P. Di Lorenzo, "Joint allocation of computation and communication resources in multiuser mobile cloud computing," in *Proc. IEEE 14th Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jun. 2013, pp. 26–30.

[11] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint optimization of radio and computational resources for multicell mobile-edge computing," *IEEE Trans. Signal Inf. Process. over Netw.*, vol. 1, no. 2, pp. 89–103, Jun. 2015.

[12] J. Kim, S.-H. Park, O. Simeone, I. Lee, and S. Shamai (Shitz), "Joint design of fronthauling and hybrid beamforming for downlink C-RAN systems," *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4423–4434, Jun. 2019.

[13] A. Liu, X. Chen, W. Yu, V. K. N. Lau, and M.-J. Zhao, "Two-timescale hybrid compression and forward for massive MIMO aided C-RAN," *IEEE Trans. Signal Process.*, vol. 67, no. 9, pp. 2484–2498, May 2019.

[14] Q. Li, J. Lei, and J. Lin, "Min-max latency optimization for multi-user computation offloading in fog-radio access networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 3754–3758.

[15] O. Dhifallah, H. Dahrouj, T. Y. Al-Naffouri, and M.-S. Alouini, "Joint hybrid backhaul and access links design in cloud-radio access networks," in *Proc. IEEE 82nd Veh. Technol. Conf. (VTC-Fall)*, Sep. 2015, pp. 1–5.

[16] E. De Carvalho, A. Ali, A. Amiri, M. Angjelichinoski, and R. W. Heath, Jr., "Non-stationarities in extra-large scale massive MIMO," 2019, *arXiv:1903.03085*. [Online]. Available: http://arxiv.org/abs/1903.03085

[17] A. Amiri, M. Angjelichinoski, E. de Carvalho, and R. W. Heath, Jr., "Extremely large aperture massive MIMO: Low complexity receiver architectures," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2018, pp. 1–6.

[18] S. Hu, F. Rusek, and O. Edfors, "Beyond massive MIMO: The potential of positioning with large intelligent surfaces," *IEEE Trans. Signal Process.*, vol. 66, no. 7, pp. 1761–1774, Apr. 2018.

[19] A. O. Martinez, E. De Carvalho, and J. O. Nielsen, "Towards very large aperture massive MIMO: A measurement based study," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2014, pp. 281–286.

[20] A. Checko *et al.*, "Cloud RAN for mobile networks—A technology overview," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 405–426, 1st Quart., 2015.

[21] A. Li and C. Masouros, "Hybrid analog-digital millimeter-wave MU-MIMO transmission with virtual path selection," *IEEE Commun. Lett.*, vol. 21, no. 2, pp. 438–441, Feb. 2017.

[22] X. Yu, J.-C. Shen, J. Zhang, and K. B. Letaief, "Alternating minimization algorithms for hybrid precoding in millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 485–500, Apr. 2016.

[23] F. Sohrabi and W. Yu, "Hybrid digital and analog beamforming design for large-scale antenna arrays," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 501–513, Apr. 2016.

[24] A. Li and C. Masouros, "Hybrid precoding and combining design for millimeter-wave multi-user MIMO based on SVD," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.

[25] W. Chen, D. Wang, and K. Li, "Multi-user multi-task computation offloading in green mobile edge cloud computing," *IEEE Trans. Services Comput.*, vol. 12, no. 5, pp. 726–738, Sep. 2019.

[26] J. Zheng, Y. Cai, Y. Wu, and X. Shen, "Dynamic computation offloading for mobile cloud computing: A stochastic game-theoretic approach," *IEEE Trans. Mobile Comput.*, vol. 18, no. 4, pp. 771–786, Apr. 2019.

[27] X. Yang, Z. Fei, J. Zheng, N. Zhang, and A. Anpalagan, "Joint multi-user computation offloading and data caching for hybrid mobile cloud/edge computing," *IEEE Trans. Veh. Technol.*, vol. 68, no. 11, pp. 11018–11030, Nov. 2019.

[28] X. Diao, J. Zheng, Y. Wu, Y. Cai, and A. Anpalagan, "Joint trajectory design, task data, and computing resource allocations for NOMA-based and UAV-assisted mobile edge computing," *IEEE Access*, vol. 7, pp. 117448–117459, 2019.

[29] R. Dong, C. She, W. Hardjawana, Y. Li, and B. Vucetic, "Deep learning for hybrid 5G services in mobile edge computing systems: Learn from a digital twin," *IEEE Trans. Wireless Commun.*, vol. 18, no. 10, pp. 4692–4707, Oct. 2019.

[30] W. Lee, M. Kim, and D.-H. Cho, "Deep power control: Transmit power control scheme based on convolutional neural network," *IEEE Commun. Lett.*, vol. 22, no. 6, pp. 1276–1279, Jun. 2018.

[31] F. Liang, C. Shen, and F. Wu, "An iterative BP-CNN architecture for channel decoding," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 1, pp. 144–159, Feb. 2018.

[32] J. Xu, P. Zhu, J. Li, and X. You, "Deep learning-based pilot design for multi-user distributed massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 8, no. 4, pp. 1016–1019, Aug. 2019.

[33] K. Kim, J. Lee, and J. Choi, "Deep learning based pilot allocation scheme (DL-PAS) for 5G massive MIMO system," *IEEE Commun. Lett.*, vol. 22, no. 4, pp. 828–831, Apr. 2018.

[34] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, no. 5, pp. 359–366, Jan. 1989.

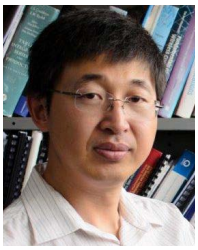
Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on April 29,2022 at 09:05:06 UTC from IEEE Xplore. Restrictions apply.

[35] H. Sun, X. Chen, Q. Shi, M. Hong, X. Fu, and N. D. Sidiropoulos, "Learning to optimize: Training deep neural networks for interference management," *IEEE Trans. Signal Process.*, vol. 66, no. 20, pp. 5438–5453, Oct. 2018.
[36] N. C. Luong *et al.*, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3133–3174, 2019.
[37] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *Proc. 27th Int. Conf. Artif. Neural Netw.*, Rhodes, Greece, Oct. 2018, pp. 270–279.
[38] A. Zappone, M. Di Renzo, and M. Debbah, "Wireless networks design in the era of deep learning: Model-based, AI-based, or both?" 2019, *arXiv:1902.02647*. [Online]. Available: http://arxiv.org/abs/1902.02647
[39] L. Liu, S. Bi, and R. Zhang, "Joint power control and fronthaul rate allocation for throughput maximization in OFDMA-based cloud radio access network," *IEEE Trans. Commun.*, vol. 63, no. 11, pp. 4097–4110, Nov. 2015.
[40] C. Zhang, P. Patras, and H. Haddadi, "Deep learning in mobile and wireless networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2224–2287, 3rd Quart., 2019.
[41] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.
[42] D. Zhang, P. Pan, R. You, and H. Wang, "SVD-based low-complexity hybrid precoding for millimeter-wave MIMO systems," *IEEE Commun. Lett.*, vol. 22, no. 10, pp. 2176–2179, Oct. 2018.
[43] J. C. Bezdek and R. J. Hathaway, "Some notes on alternating optimization," in *Proc. AFSS Int. Conf. Fuzzy Syst.*, Calcutta, India, Feb. 2002, pp. 288–300.
[44] B. R. Marks and G. P. Wright, "A general inner approximation algorithm for non-convex mathematical programs," *Oper. Res.*, vol. 26, no. 2, pp. 681–683, Aug. 1978.
[45] A. Zappone, L. Sanguinetti, G. Bacci, E. Jorswieck, and M. Debbah, "Energy-efficient power control: A look at 5G wireless technologies," *IEEE Trans. Signal Process.*, vol. 64, no. 7, pp. 1668–1683, Apr. 2016.
[46] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[47] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html
[48] F. Chollet *et al.* (2015). *Keras*. [Online]. Available: https://keras.io
[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[50] Y. Shen, Y. Shi, J. Zhang, and K. B. Letaief, "Transfer learning for mixed-integer resource allocation problems in wireless networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–6.
[51] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[52] R. D. Yates, "A framework for uplink power control in cellular radio systems," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 7, pp. 1341–1347, Sep. 1995.

Chandan Pradhan (Student Member, IEEE) received the B.Tech. degree from IIIT Bhubaneswar, India, in 2013, and the and M.S. degree from IIIT Hyderabad, India, in 2016. He is currently pursuing the Ph.D. degree with the School of Electrical and Information Engineering, The University of Sydney, Sydney, Australia. His research interests lie in the fields of beamforming techniques in MIMO systems and application of deep-learning in wireless networks.

Ang Li (Member, IEEE) received the Ph.D. degree from the Communications and Information Systems Research Group, Department of Electrical and Electronic Engineering, University College London, U.K., in 2018. He was a Post-Doctoral Research Associate with the School of Electrical and Information Engineering, The University of Sydney, Sydney, Australia, from May 2018 to February 2020. He is currently with the Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China. His research interests lie in the field of beamforming and signal processing techniques for MIMO systems. He was a recipient of the Exemplary Reviewer for IEEE COMMUNICATIONS LETTERS and IEEE TRANSACTIONS ON COMMUNICATIONS in 2017 and 2019.

Changyang She (Member, IEEE) received the B.Eng. degree from the Honors College (formerly the School of Advanced Engineering), Beihang University (BUAA), Beijing, China, in 2012, and the Ph.D. degree from the School of Electronics and Information Engineering, BUAA, in 2017. From 2017 to 2018, he was a Post-Doctoral Research Fellow with the Singapore University of Technology and Design. Since 2018, he has been a Post-Doctoral Research Associate with the University of Sydney. His research interests lie in the areas of ultra-reliable and low-latency communications, deep learning in wireless networks, mobile edge computing, and energy efficient 5G communication systems.



Yonghui Li (Fellow, IEEE) received the Ph.D. degree from the Beijing University of Aeronautics and Astronautics, in November 2002. From 1999 to 2003, he was affiliated with Linkair Communication Inc., where he held a position of the Project Manager with responsibility for the design of physical layer solutions for the LAS-CDMA system. Since 2003, he has been with the Centre of Excellence in Telecommunications, The University of Sydney, Australia, where he is currently a Professor with the School of Electrical and Information Engineering. His current research interests are in the area of wireless communications, with a particular focus on MIMO, millimeter wave communications, machine to machine communications, coding techniques, and cooperative communications. He holds a number of patents granted and pending in these fields. He was a recipient of the Australian Queen Elizabeth II Fellowship in 2008 and the Australian Future Fellowship in 2012. He received the best paper awards from IEEE International Conference on Communications (ICC) 2014, IEEE Wireless Days Conferences (WD) 2014, and IEEE PIMRC 2017. He has also served as a Guest Editor for several special issues of IEEE Journals, such as IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS Special Issue on Millimeter Wave Communications. He is also an Editor of IEEE TRANSACTIONS ON COMMUNICATIONS and IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY.

Branka Vucetic (Life Fellow, IEEE) is currently an ARC Laureate Fellow and the Director of the Centre of Excellence for Internet of Things (IoT) and Telecommunications, The University of Sydney. Her current research work is in wireless networks and the IoT. In the area of wireless networks, she works on ultra-reliable low-latency communications (URLLC) and system design for millimetre wave frequency bands. In the area of the IoT, Vucetic works on providing wireless connectivity for mission critical applications. She is a Fellow of the Australian Academy of Technological Sciences and Engineering and the Australian Academy of Science.

Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on April 29,2022 at 09:05:06 UTC from IEEE Xplore. Restrictions apply.