IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 November 2022, accepted 15 December 2022, date of publication 16 December 2022,
date of current version 27 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3230326

RESEARCH ARTICLE

An Efficient Beamforming Architecture to Handle the Trade-Off Between Performance and Hardware Complexity in Multiuser Massive MISO Systems

JAMAL BEIRANVAND, VAHID MEGHDADI, CYRILLE MENUDIER, (Member, IEEE), AND JEAN PIERRE CANCES, (Member, IEEE)
XLIM UMR 7252 Laboratory, University of Limoges, 87000 Limoges, France
Corresponding author: Jamal Beiranvand (jamal.beiranvand@etu.unilim.fr)

ABSTRACT Beamforming is the fundamental concept of wireless communications to serve several users through multiple-antenna transceivers. The advent of massive multiple-input multiple-output (MIMO) systems leads to an investigation into beamforming to reach optimal performance. However, fully-digital beamforming implementation has to face important challenges in terms of consumed power and cost of the RF chains and converters. To solve the problem, some studies focus on reducing hardware complexity by dividing the beamforming into digital and analog parts, known as hybrid beamforming (HBF). In HBF systems, the dominant part of the hardware complexity depends on the architecture of the analog network, which determines the required RF components. In this paper, we categorize the analog beamforming parts based on the connectivity between RF-chains and antennas. Moreover, we show the impacts of the connection on the analog beamforming matrix. To this aim, we propose an efficient connectivity architecture in which the antennas are divided into fully-connected and singly-connected groups so that we can deal with the hardware complexity-performance trade-off. In addition, to achieve further performance improvements, we propose a dynamic architecture that adjusts the connection of the RF-chain antenna to the channel state information (CSI) through a switch network. Analytically, we propose an efficient approach to calculate the zero-forcing precoder, which is proper for both fixed and dynamic architectures. A two-part algorithm based on the greedy search method has also been developed to obtain switch states in dynamic architecture and then implemented by a deep neural network (DNN). The simulation results confirm the theoretical analysis and the suitability of the proposed architecture.

INDEX TERMS Millimeter wave communications, massive MIMO, hybrid beamforming, fixed phase shifter.

I. INTRODUCTION

Thus far, the wireless networks have achieved a capacity increase appropriate to the data traffic demands due to the enhancements in the physical layer [1], [2], [3]. Since the techniques at the physical layer are insufficient to achieve more efficiency improvement, exploring less-congested spectrum bands is unavoidable to meet further data traffic demands [4]. Large free bandwidths in millimeter wave (mmWave) band and developments in mmWave-hardware devices encourage the wireless industry to design mmWave systems for the fifth generation (5G) and beyond to increase the network capacity [5], [6], [7]. Path loss, as an inherent effect of mmWave signals, degrades the spectral efficiency of the wireless communication link. However, a redeeming property in mmWave signals is a decrease in wavelength that allows occupying of a large number of antenna elements in a small space [7], [8], [9]. Therefore,

The associate editor coordinating the review of this manuscript and approving it for publication was Stefan Schwarz.

it is possible to densify antenna arrays in cities by keeping a discreet visual footprint. It is then possible to generate high gain with beamforming capabilities in mmWave systems even with the path loss effects. Furthermore, this enables the wireless systems to approach the objective capacity, potentially by precoding multiple data streams [10], [11].

While conventional precoding techniques for microwave systems are performed digitally at the baseband, which enables accessing both the signal phase and the amplitude; the mmWave imposes crucial practical constraints on signal processing in mmWave MIMO systems. Particularly, the cost and power consumption of mmWave radio frequency (RF) chain (analog-to-digital converters (ADC)/digital-to-analog converters (DAC), data converters, mixers, RF amplifiers, etc.) force the industry not to dedicate an RF-chain per MIMO antenna element as in microwave systems [12], [13], [14]. Hence, the signal processing in mmWave systems relies on RF processing that is under constraints due to its implementation via RF components such as phase shifters and switches. Several works investigate two-part precoding approaches, known as hybrid beamforming (HBF), in such mmWave MIMO systems [13], [14], [15], [16], [17], [18], [19]. Regarding hardware complexity, the RF segment, implementing the RF precoder, is the dominant part that relies on two factors: the *RF components* and the *RF architecture*. The RF components specify the feasible set for selecting the analog beamforming coefficients, and the RF architecture determines the *RF-paths* connecting the RF-chains to the antenna elements. Most published studies have considered traditional architectures, called the *fully connected* (FC) and the *partially connected* (PC), to develop the hybrid precoding approaches [15], [16], [17], [18], [19], [20], [21], [22], [23]. The FC-HBF effectively benefits the analog precoding gain by transmitting a combination of RF signals through the antenna elements. Nevertheless, the drawback of this system is a large number of RF components that cause an increase in power consumption and cost. On the other hand, the PC-HBF reduces the required number of RF components by restricting the antenna elements to be connected to only one RF-chain. Although the PC-HBF reduces the hardware complexity, power consumption, and cost of implementation, it suffers from severe degradation of the performance. The system design trade-off, between performance and hardware complexity, has attracted the attention of academia and industry [13].

Regardless the RF components, the performance and the complexity are strongly linked to RF architecturesEspecially, more RF-paths yield to better performance but with a higher complexity. In the light of this, two architectures have been suggested, the *group connected* (GC) and the *overlapped subarray* (OSA), in order to balance the performance and the hardware complexity ratio [18], [24]. The GC separates the antenna elements and the RF-chains into disjoint groups, up to the number of RF-chains, with the FC strategy applied for the connection into the groups. While the number of subarrays (groups) in the OSA is fixed and equal to the RF-chains, the subarrays are allowed to overlap. The GC and OSA architectures enable us to save the hardware complexity by adjusting the number of groups and overlapped antenna elements. Nonetheless, the limited number of implementable architectures in both approaches implies substantial gaps between the different performance-hardware levels in those systems.

Inspired by the aforementioned architectures, adaptive architectures have been put forward to attain a significant performance enhancement by adding RF switches to adjust the RF-paths (from RF-chains to antennas) or enable/disable RF components [25], [26], [27]. The authors in [26] have developed a dynamic PC-HBF by allotting a switch between each RF-chain and subarray. The switch states are adjusted to minimize the power consumption for a given data rate. This architecture has a phase shifter in each RF-path. Thus, it becomes the FC when all switches are on. In [27], a switch is dedicated to a phase shifter (or RF-path) in which the switch states are obtained based on the maximum energy efficiency (EE) criterion. Dynamic architectures deploy switch networks for two main reasons: the EE and the spectral efficiency (SE). Higher EEs are achieved by reducing energy consumption resulting from *disabling* specific RF-paths (a phase shifter, an element/subarray of antenna array), and an enhancement in the SE is obtained by *adjusting* the RF-paths to the CSI. Note that a dynamic architecture imposes additional hardware requirements (switch network) compared to the corresponding fixed architecture.

Motivated by the above discussion, this paper investigates architecture design and develops precoder algorithms for multi-user massive multiple-input single-output (MISO) systems to settle the dichotomy between performance and hardware complexity. The major contributions are:

1) A novel architecture, called the partially/fully-connected (PFC) architecture, is proposed for massive MISO systems. It can control the hardware complexity through an insight into the effects of RF-paths connecting RF-chains to antennas. In this architecture, the antenna elements are divided into two groups: the fully-connected antennas (FCAs) and the singly-connected antennas (SCAs), where the FCAs scale up the number of RF-paths by a factor linked to the number of RF-chains. Therefore, the system design strongly depends on the number of FCAs, which allows us to develop systems at different performance-complexity ratios. Furthermore, the proposed algorithm, canceling user interference in this architecture, eliminates zero-entry elements from the precoder matrix and reforms the optimization problem decreasing the complexity burden.
2) To further improve the performance, a dynamic architecture adapts RF-paths (selecting FCAs) to the CSI by inserting a switch network between the RF-chains and the antenna array. It poses the problem of the antenna selection; i.e., selecting the FCAs and allocating the SCAs to the RF-chains. Since exhaustive search is impractical, especially in massive arrays,

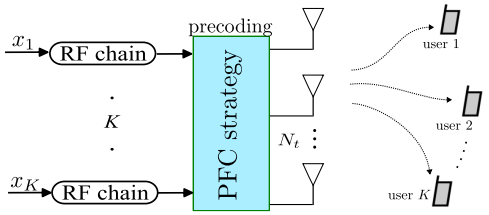

FIGURE 1. The analog BF designed by the proposed PFC architecture.

a two-part approach is proposed. First, FCAs are iteratively selected through a greedy algorithm. Next, the second algorithm assigns the remaining SCAs to the users. A deep neural network that is trained by the algorithm outputs and the CSI is developed to perform the approach in real-time systems. Furthermore, post-processing is applied to satisfy the hardware limitations.

Organization: the paper is organized as follows:

Section II introduces the PFC architecture and corresponding zero-forcing precoder in massive MISO systems. Section III develops the dynamic PFC architecture and antenna assignment algorithms. The proposed DNN for the dynamic PFC architecture is presented in Section IV. Simulation results are presented in Section V, and section VI concludes the paper.

Notations: the following notations are used throughout the present paper: boldface upper-case letters and boldface lower-case letters denote matrices and vectors, respectively. $\mathbf{A}(i, \ell)$ denotes the entry in ith row and ℓth column of matrix $\mathbf{A}$, $(.)^{\mathrm{T}}$ and $(.)^{\mathrm{H}}$ denote the transpose and conjugate transpose operations, respectively. $\|.\|_{\mathrm{F}}^2$ denotes Frobenius norm, the notations $\mathbb{R}$ and $\mathbb{C}$ stand for the sets of real and complex numbers, respectively. $\mathcal{O}(.)$ shows how the run time grows with the input size. The mathematical expectation operator is represented by $\mathbb{E}\{.\}$, $\mathcal{CN}\left(0, \sigma^2\right)$ represents a zero mean complex Gaussian distribution with variance σ^2, the $N \times N$ identity matrix is represented by $\mathbf{I}_N$, $\lfloor . \rfloor$ is the floor function, and $[N] = \{1, 2, \ldots, N\}$.

II. SYSTEM MODEL

Let us consider the downlink transmission of a multi-user massive MISO system, where a base station (BS) is equipped with an antenna array of size N_t to serve K single-antenna users as depicted in Figure 1. The BS is designed based on the PFC strategy, as explained in subsection A, to perform the analog beamforming. It is shown that we can achieve the optimal performance of the fully digital BF for flat fading channels by using fully analog BF [21], [23], [28]. Accordingly, in this paper, the digital BF part is not used and the number of RF-chains, N_{RF}, is considered equal to the minimum, i.e., the number of data streams, K. The BS applies the precoder matrix $\mathbf{F} \in \mathbb{C}^{N_t \times K}$ on data stream vector $\mathbf{x}$, where $\mathbb{E}\{\mathbf{x}\mathbf{x}^{\mathrm{H}}\} = \mathbf{I}_K$. Therefore, the received signals on the users' side are given by:

$$\mathbf{y} = \sqrt{\rho}\mathbf{H}\mathbf{F}\mathbf{x} + \mathbf{n}, \tag{1}$$

where $\mathbf{H} \in \mathbb{C}^{K \times N_t}$ represents the channel matrix, $\mathbf{n} \in \mathbb{C}^{K \times 1}$ is the complex Gaussian noise vector with $\mathbf{n} \sim \mathcal{CN}(0, \mathbf{I}_K)$, and the normalization constant ρ is chosen to respect the transmit power constraint P:

$$\rho = \frac{P}{\|\mathbf{F}\|_{\mathrm{F}}^2}. \tag{2}$$

In the following subsection, the proposed PFC architecture is presented, which is capable of handling the trade-off between hardware complexity and performance.

A. PARTIALLY/FULLY-CONNECTED (PFC) ARCHITECTURE

Analog BF systems can be categorized, in general, based on the connection strategy in the analog part, as illustrated in Figure 2. As shown in Figure 2(a), the FC strategy connects all the RF-chains to all the antennas through $N_{RF}N_t$ RF-paths. Although this architecture requires the most complex hardware, it puts no constraint on the precoder matrix, allowing the achievement of the optimal performance. Notice that each entry in the precoding matrix $\mathbf{F}$ represents an RF-path connecting an RF-chain to an antenna.

On the other hand, the PC architecture, the simplest one, has one RF-path per antenna that drastically simplifies the hardware. It results in a block-diagonal precoder matrix with only N_t non-zero entries. In fact, a large number of zeros in the matrix represent the deleted connections compared with the FC strategy, which tumbles the performance.

The GC architecture, depicted in Figure 2(c), divides RF-chains and antennas into disjoint groups, which deploy the FC strategy for inside-group connections. The complexity of the architecture lies between the FC and PC. As a result, more RF-paths exist in the system, which means there are more non-zero elements in the matrix. A certain flexibility is obtained because the number of groups can vary from 1 (equivalent to the FC architecture) to N_{RF} (equivalent to the PC). Therefore, the number of non-zero elements (or RF-paths) to be optimized can be selected from the set $\{N_t, 2N_t, \ldots, N_{RF}N_t\}$.

As shown in Figure 2(d), the OSA structure allows the groups to overlap to cope with the hardware complexity-performance challenge. Here, each RF-chain is connected to a subarray of the size $M_t = N_t - (N_{RF} - 1)\Delta M_t$, where ΔM_t represents the number of overlapped elements. Therefore, $N_t/N_{RF} + 1$ different levels of performance can be designed by adjusting the number of overlapped elements ΔM_t. In fact, FC and PC architectures are two special cases for $\Delta M_t = 0$ and $\Delta M_t = N_t/N_{RF}$, respectively. The GC and OSA, nevertheless, offer the view of adjusting the number of RF-paths, and it seems that the development of this approach is essential for the effective management of the hardware performance trade-off.

In this paper, we propose a novel connectivity that we call partially/fully-connected (PFC). It is capable of adjusting the number of RF-paths in $N_t + 1$ different levels, while the precoder matrix still has a simple form, suited to analytical optimization. As shown in Figure 2(e), we divide the available

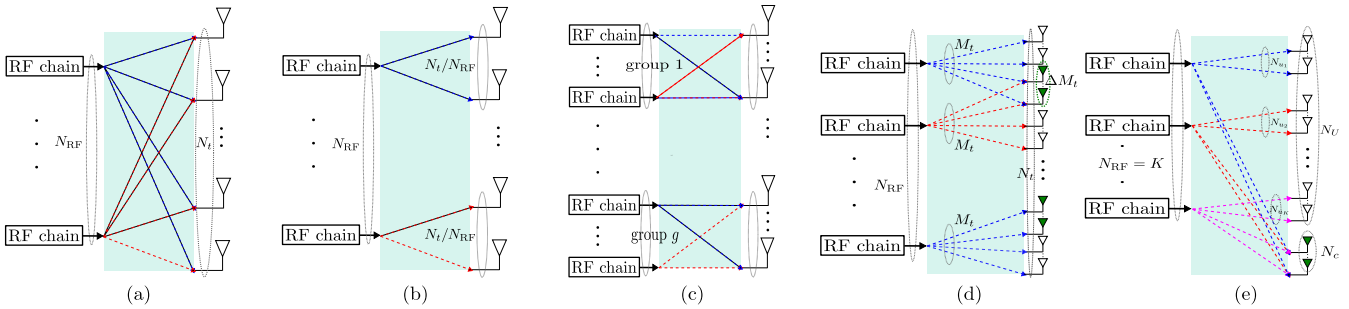

FIGURE 2. The analog BF connection architectures: (a) the FC strategy, (b) the PC strategy, (c) the GC strategy, (d) the OSA strategy with ΔM_t overlapped antennas, (e) the proposed PFC strategy with N_c FCAs.

antennas into two groups. The first group, comprising N_U antennas, is implemented as PC, i.e. one RF-path per antenna. The second group, comprising the remaining antennas, $N_c = N_t - N_U$, is implemented via FC architecture, i.e. each antenna is connected to all the RF-chains. We call the antennas of the first group singly-connected antennas (SCAs) and those of the second group fully-connected-antennas (FCAs). We note that, $N_c = 0$ converges to the PC strategy and $N_c = N_t$ to the FC strategy. One of the advantages of this structure is that it allows controlling the user priority by assigning more or less SCAs to a specific user, which creates flexibility in user priorities. If the number of antennas assigned to the kth user is denoted by N_{u_k}, the following expression should be verified: $\sum N_{u_k} = N_U$.

Without loss of generality, we assume the following ordering for antenna connections:

The N_{u_1} first antennas are connected to the first RF-chain, the N_{u_2} following antennas to the second RF-chain, and so on. The last N_c antennas are the FCAs connected to all the RF-chains. With this assumption, the analog precoding matrix $\mathbf{F}$ can be expressed as:

$$\mathbf{F} = \begin{bmatrix} \mathbf{f}_{u_1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{f}_{u_2} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{f}_{u_K} \\ \mathbf{f}_{c_1} & \mathbf{f}_{c_2} & \dots & \mathbf{f}_{c_K} \end{bmatrix}, \tag{3}$$

where $\mathbf{f}_{u_k} \in \mathbb{C}^{N_{u_k} \times 1}$ and $\mathbf{f}_{c_k} \in \mathbb{C}^{N_c \times 1}$ contain the beamforming coefficients of the SCAs and FCAs related to the kth RF-chain, respectively.

B. PRECODER OPTIMIZATION

In this section, we aim to obtain $\mathbf{F}$ by solving the optimization problem of the zero-forcing precoder. We do not impose any constraint on the $\mathbf{F}$ entries, thanks to the analog front-end design given in [21] and [23]. The only constraint comes from the structure simplifications obtained by deleting some RF-paths, which results in the special form of the matrix $\mathbf{F}$ as given in (3). In other words, a large number of entries must be forced to zero. Since the inter-user interference intensely curtails the performance of the multi-user systems, we design $\mathbf{F}$ in such a way that nullifies the interference. More precisely, the kth column of the precoder matrix $\mathbf{F}$, denoted by $\mathbf{f}_k$, shall satisfy:

$$\begin{cases} \mathbf{h}_k \mathbf{f}_{k'} = 1, & k = k' \\ \mathbf{h}_k \mathbf{f}_{k'} = 0, & k \neq k'. \end{cases} \tag{4}$$

where $\mathbf{h}_k$ is the kth row of the channel matrix $\mathbf{H}$. Therefore, we form the optimization problem as:

$$\begin{aligned} \mathbf{F} = \underset{\mathbf{F}}{\text{argmin}} \quad & \|\mathbf{F}\|_{\mathrm{F}}^2 \\ \text{s.t.} \quad & \mathbf{HF} = \mathbf{I}_K. \end{aligned} \tag{5}$$

It means that from all the possible solutions, the one with minimum transmit power is selected. Since the precoder matrix $\mathbf{F}$ is a sparse matrix containing $(K-1)N_U$ zero elements, it hinders the problem (5) to be solved through classical approaches. Therefore, we place all non-zero elements, N_{nz}, of $\mathbf{F}$ in the vector $\hat{\mathbf{f}} \in \mathbb{C}^{N_{nz} \times 1}$ as:

$$\hat{\mathbf{f}} = \left[\hat{\mathbf{f}}_1, \dots, \hat{\mathbf{f}}_K\right]^{\mathrm{T}}, \tag{6}$$

in which $\hat{\mathbf{f}}_k = \left[\mathbf{f}_{u_k}^{\mathrm{T}}, \mathbf{f}_{c_k}^{\mathrm{T}}\right]$. Also, we express the channel matrix as:

$$\mathbf{H} = \left[\mathbf{H}_{\mathcal{U}_1}, \mathbf{H}_{\mathcal{U}_2}, \dots, \mathbf{H}_{\mathcal{U}_K}, \mathbf{H}_{\mathcal{C}}\right]_{K \times N_t}, \tag{7}$$

where $\mathbf{H}_{\mathcal{C}} \in \mathbb{C}^{K \times N_c}$ is the sub-matrix of the channel matrix corresponding to the FCAs, and $\mathbf{H}_{\mathcal{U}_k} \in \mathbb{C}^{K \times N_{u_k}}$ is the channel sub-matrix corresponding to the N_{u_k} SCAs assigned to the kth user. Now, (5) can be rewritten as:

$$\begin{aligned} \hat{\mathbf{f}} = \underset{\hat{\mathbf{f}}}{\text{argmin}} \quad & \left\|\hat{\mathbf{f}}\right\|_{\mathrm{F}}^2 \\ \text{s.t.} \quad & \hat{\mathbf{H}}\hat{\mathbf{f}} = \hat{\mathbf{i}}, \end{aligned} \tag{8}$$

where $\hat{\mathbf{i}} = \mathrm{vec}(\mathbf{I})$ and $\hat{\mathbf{H}} \in \mathbb{C}^{K^2 \times N_{nz}}$ is defined as:

$$\hat{\mathbf{H}} = \mathrm{Blckdg}(\mathbf{H}_1, \dots, \mathbf{H}_K), \tag{9}$$

and $\mathbf{H}_k = \left[\mathbf{H}_{\mathcal{U}_k} | \mathbf{H}_{\mathcal{C}}\right]$. The problem (8), under some conditions discussed below, has a straightforward solution as:

$$\hat{\mathbf{f}} = \hat{\mathbf{H}}^{\mathrm{H}} \left(\hat{\mathbf{H}}\hat{\mathbf{H}}^{\mathrm{H}}\right)^{-1} \hat{\mathbf{i}}. \tag{10}$$

Once $\hat{\mathbf{f}}$ is computed, we put its entries back to the matrix $\mathbf{F}$ by using (6).

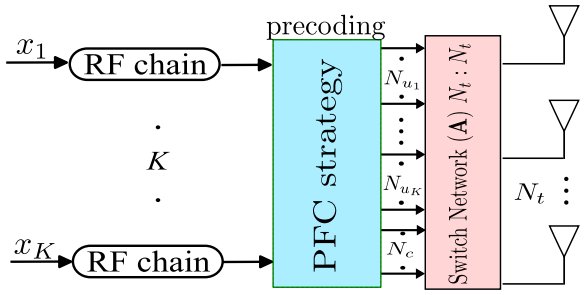

FIGURE 3. The BS architecture deploying the dynamic PFC strategy.

C. DISCUSSIONS

1) Since the matrix $\hat{\mathbf{H}}$ is a block diagonal matrix, $\hat{\mathbf{f}}_k$ can be computed independently of the others as:

$$\hat{\mathbf{f}}_k = \mathbf{H}_k^{\mathrm{H}} \left(\mathbf{H}_k \mathbf{H}_k^{\mathrm{H}}\right)^{-1} \mathbf{i}_k, \tag{11}$$

where $\mathbf{i}_k$ is the kth column of $\mathbf{I}_K$. This observation will greatly reduce the computation complexity.

2) In order for $\mathbf{H}_k \mathbf{H}_k^{\mathrm{H}}$ to be an invertible matrix, the following constraint is required: $N_{u_k} + N_c \geq K, \forall k$.

3) In the case of fully-connected strategy, where $N_c = N_t$, we have:

$$\mathbf{H}_k = \mathbf{H} \qquad \forall k = 1, \ldots, K. \tag{12}$$

By substituting (12) in (11), $\hat{\mathbf{f}}_k = \mathbf{H}^{\mathrm{H}} \left(\mathbf{H}\mathbf{H}^{\mathrm{H}}\right)^{-1} \mathbf{i}_k$, where $\hat{\mathbf{f}}_k$ is exactly the kth column of the $\mathbf{F}$, therefore:

$$\mathbf{F} = \mathbf{H}^{\mathrm{H}} \left(\mathbf{H}\mathbf{H}^{\mathrm{H}}\right)^{-1}, \tag{13}$$

which is the conventional zero-forcing precoder.

III. DYNAMIC PFC ARCHITECTURE

In the previous section, we did not focus on how to select the FCAs and how to assign the SCAs to the users. Obviously, the optimal selection of antennas depends on CSI to achieve SE improvement. Here, given the channel, we propose an antenna assignment method that adapts the RF-paths to CSI via a switch network controlling the connection of the outputs to the antennas dynamically. It is named ''dynamic PFC architecture''. Mathematically, it can be obtained by the multiplication of the outputs by a permutation matrix $\mathbf{A}$ of size $N_t \times N_t$. Therefore, received signals at the users' side can be written as:

$$\mathbf{y} = \sqrt{\rho}\mathbf{H}\mathbf{A}\mathbf{F}\mathbf{x} + \mathbf{n}. \tag{14}$$

Let us define the set $\mathcal{C}$ containing the indexes of FCAs and the set $\mathcal{U}_k$ containing the indexes of SCAs assigned to the kth user. The switch matrix $\mathbf{A}$ can be written as:

$$\mathbf{A} = [\mathbf{a}_{\mathcal{P}(1)}, \ldots, \mathbf{a}_{\mathcal{P}(N_t)}], \tag{15}$$

where $\mathbf{a}_{\mathcal{P}(i)}$ denotes a column vector of length N_t with 1 in the $\mathcal{P}(i)$th position and 0 elsewhere, and the permutation $\mathcal{P}$ is defined as:

$$\mathcal{P} : \{1, \ldots, N_t\} \rightarrow \{\mathcal{U}_1, \ldots, \mathcal{U}_K, \mathcal{C}\}. \tag{16}$$

By taking into account the permutation, we define the new channel matrix as:

$$\begin{aligned} \widetilde{\mathbf{H}} &= \mathbf{H}\mathbf{A} \\ &= \left[\mathbf{H}_{\mathcal{U}_1}, \mathbf{H}_{\mathcal{U}_2}, \ldots, \mathbf{H}_{\mathcal{U}_K}, \mathbf{H}_{\mathcal{C}}\right]_{K \times N_t}, \end{aligned} \tag{17}$$

where $\mathbf{H}_{\mathcal{U}_k} \in \mathbb{C}^{K \times N_{u_k}}$ and $\mathbf{H}_{\mathcal{C}} \in \mathbb{C}^{K \times N_c}$ consist of the columns indicated in $\mathcal{U}_k$ and $\mathcal{C}$, respectively. If the switch matrix $\mathbf{A}$ is known, the beamforming matrix $\mathbf{F}$ is derived as before from (5) by substituting $\mathbf{H}$ by $\widetilde{\mathbf{H}}$.

Since $\mathbf{H}\mathbf{A}\mathbf{F} = \mathbf{I}_K$, as before, we minimize the $\|\mathbf{F}\|_{\mathrm{F}}^2$ but for all the possible combinations given by $\mathbf{A}$. Therefore, the optimal permutation can be obtained from:

$$\begin{aligned} \left(\mathcal{C}^{opt}, \mathcal{U}_1^{opt}, \ldots, \mathcal{U}_K^{opt}\right) = \underset{\mathcal{C}, \mathcal{U}_1, \ldots, \mathcal{U}_K}{\operatorname{argmin}} \quad & \|\mathbf{F}\|_{\mathrm{F}}^2 \\ \text{s.t.} \quad & \mathbf{H}\mathbf{A}\mathbf{F} = \mathbf{I}_K. \end{aligned} \tag{18}$$

Since $\|\mathbf{F}\|_{\mathrm{F}}^2 = \sum_{k=1}^{K} \left\|\hat{\mathbf{f}}_k\right\|^2$, we can simplify the argument of the summation as:

$$\begin{aligned} \left\|\hat{\mathbf{f}}_k\right\|_{\mathrm{F}}^2 &= \hat{\mathbf{f}}_k^{\mathrm{H}} \hat{\mathbf{f}}_k \\ &= \mathbf{i}_k^{\mathrm{H}} \left(\left(\mathbf{H}_k \mathbf{H}_k^{\mathrm{H}}\right)^{-1}\right)^{\mathrm{H}} \mathbf{H}_k \mathbf{H}_k^{\mathrm{H}} \left(\mathbf{H}_k \mathbf{H}_k^{\mathrm{H}}\right)^{-1} \mathbf{i}_k \\ &= \mathbf{i}_k^{\mathrm{H}} \left(\left(\mathbf{H}_k \mathbf{H}_k^{\mathrm{H}}\right)^{-1}\right)^{\mathrm{H}} \mathbf{i}_k \\ &= \left[\left(\mathbf{H}_k \mathbf{H}_k^{\mathrm{H}}\right)^{-1}\right]_{k,k}. \end{aligned} \tag{19}$$

By substituting (19) in (18), we have:

$$\left(\mathcal{C}^{opt}, \mathcal{U}_1^{opt}, \ldots, \mathcal{U}_K^{opt}\right) = \underset{\mathcal{C}, \mathcal{U}_1, \ldots, \mathcal{U}_K}{\operatorname{argmin}} \quad \sum_{k=1}^{K} \left[\left(\mathbf{H}_k \mathbf{H}_k^{\mathrm{H}}\right)^{-1}\right]_{k,k}. \tag{20}$$

An advantage is that the computation of the optimized $\mathbf{F}$ is not required at each permutation, only $\|\mathbf{F}\|_{\mathrm{F}}^2$ is needed, using the last simplified expression. However, the number of permutations is prohibitive to make an exhaustive search. In the following section, we propose a heuristic method with two relatively simple algorithms: the first one selects the FCAs, and the second one assigns the SCAs to users.

A. ALGORITHM 1: SELECTING FCAs

Since most of the beamforming is done by common antennas, we prioritize the FCAs selection. Intuitively, we select the most important antenna, N_c columns of the matrix $\mathbf{H}$, which is the most informative part of $\mathbf{H}$. This problem has been addressed in [29], according to which, we face the following optimization problem:

$$\mathcal{C} = \underset{\mathcal{C}}{\operatorname{argmin}} \quad \mathrm{Tr}\left\{\left(\mathbf{H}_{\mathcal{C}} \mathbf{H}_{\mathcal{C}}^{\mathrm{H}}\right)^{-1}\right\}, \tag{21}$$

where the set $\mathcal{C}$ contains the indexes of the selected columns. To solve this complex combinatorial column selection problem, [30] proposes a less complex greedy removal method.

Algorithm 1 selecting FCAs

Input: $\mathbf{H}$, N_c
$\mathcal{C}^{(0)} = \{\ell \in \mathbb{N} : \ell \leq N_t\}$
For $t = 1, \ldots, N_t - N_c$
  $\ell = \text{argmin}_\ell \quad \text{Tr}\left\{\left(\mathbf{H}_{\mathcal{C}^{(t)}}\mathbf{H}_{\mathcal{C}^{(t)}}^{\text{H}} - \mathbf{h}_\ell \mathbf{h}_\ell^{\text{H}}\right)^{-1}\right\}$
  s.t. $\quad \ell \in \mathcal{C}^{(t)}$
  $\mathcal{C}^{(t+1)} = \mathcal{C}^{(t)} - \{\ell\}$
end
Output: $\mathcal{C}$ and $\mathbf{H}_{\mathcal{C}}$

Therefore, we iteratively identify the least informative column and eliminate it from the matrix. The algorithm starts with $\mathcal{C}^{(0)} = [N_t]$. Then, one member is removed in each iteration. For instance, in the tth iteration of the algorithm, $\ell \in \mathcal{C}^{(t)}$ is found from:

$$\ell = \underset{\ell}{\text{argmin}} \quad \text{Tr}\left\{\left(\mathbf{H}_{\mathcal{C}^{(t)}}\mathbf{H}_{\mathcal{C}^{(t)}}^{\text{H}} - \mathbf{h}_\ell \mathbf{h}_\ell^{\text{H}}\right)^{-1}\right\}, \quad s.t. \;\; \ell \in \mathcal{C}^{(t)} \tag{22}$$

then ℓ is removed for the next iteration, i.e, $\mathcal{C}^{(t+1)} = \mathcal{C}^{(t)} - \ell$, and the ℓth column of the matrix $\mathbf{H}_{\mathcal{C}}^{(t)}$ to give $\mathbf{H}_{\mathcal{C}}^{(t+1)}$. The algorithm proceeds by removing one element at a time until $|\mathcal{C}^{(t)}| = N_c$. A complete pseudocode of the algorithm is given in the Algorithm 1.

B. ALGORITHM 2: ALLOCATING SCAs TO USERS

The second algorithm aims to assign the remaining antennas, in the set $\mathcal{N} = [N_t] - \mathcal{C}$, to the users in such a way that the Frobenius norm of $\mathbf{F}$ is minimized. For a given $\mathcal{C}$, the problem (20) can be written as:

$$(\mathcal{U}_1, \ldots, \mathcal{U}_K) = \underset{\mathcal{U}_1,\ldots,\mathcal{U}_K}{\text{argmin}} \quad \sum_{k=1}^{K}\left[\left(\mathbf{H}_k\mathbf{H}_k^{\text{H}}\right)^{-1}\right]_{k,k}. \tag{23}$$

As shown in Algorithm 2, we start with the empty sets $\mathcal{U}_k$ $(k = 1, \ldots, K)$ and add iteratively selected antennas to them. We construct the corresponding $\mathbf{H}_k = \mathbf{H}_{\mathcal{C}}$, for all k, then we append the carefully-selected ith column of the channel matrix, $\mathbf{h}_i$, to $\mathbf{H}_k$ as:

$$\widehat{\mathbf{H}}_{ki} = [\mathbf{h}_i, \mathbf{H}_k] \qquad i \in \mathcal{N}. \tag{24}$$

To analyze the effect of adding the ith column to the set $\mathcal{U}_k$ on the $\|\mathbf{F}\|_{\text{F}}^2$, we define the matrix Δ as:

$$\Delta(k, i) = p_k - p_{ki}, \tag{25}$$

where

$$p_k = \left[\left(\mathbf{H}_k\mathbf{H}_k^{\text{H}}\right)^{-1}\right]_{k,k}, \tag{26}$$

$$p_{ki} = \left[\left(\widehat{\mathbf{H}}_{ki}\widehat{\mathbf{H}}_{ki}^{\text{H}}\right)^{-1}\right]_{k,k}. \tag{27}$$

$\Delta(k, i)$ is interpreted as the amount of decreased power when the ith column is assigned to $\mathcal{U}_k$. Therefore, we use Δ as the metric for column selection at each iteration in such a way that the pair (ℓ, k), generating the maximum value in Δ, is selected:

$$(k, \ell) = \underset{k,\ell\in\mathcal{N}}{\text{argmax}} \quad \Delta(k, \ell) \quad \text{s.t} \quad |\mathcal{U}_k| < N_{u_k}. \tag{28}$$

The selected column ℓ is added to $\mathcal{U}_k$, which changes $\mathbf{H}_k$ and p_k. Therefore, the kth column of Δ is updated before allocating another antenna. Then, algorithm 2 proceeds until all antennas are allocated.

Algorithm 2 Allocating SCAs to the Users

Input: $\mathbf{H}$, $\mathcal{C}$;
1: Initialization:
  $\mathcal{N} = [N_t] - \mathcal{C}$;
  $\mathcal{U}_k = \{\}, \qquad \forall k = 1, \ldots, K$;
  $\mathbf{H}_k = \mathbf{H}_{\mathcal{C}}, \qquad \forall k = 1, \ldots, K$;

2: **For** $k = 1 : K$
3: $\widehat{\mathbf{H}}_{ki} = [\mathbf{h}_i, \mathbf{H}_k] \qquad \forall i \in \mathcal{N}$;
4: $p_k = \left[\left(\mathbf{H}_k\mathbf{H}_k^{\text{H}}\right)^{-1}\right]_{k,k}$;
5: $p_{ki} = \left[\left(\widehat{\mathbf{H}}_{ki}\widehat{\mathbf{H}}_{ki}^{\text{H}}\right)^{-1}\right]_{k,k}$;
6: $\Delta(k, i) = p_k - p_{ki}$;
7: **end**

8: **Repeat**
9: $(k, \ell) = \text{argmax}_{k,\ell\in\mathcal{N}} \quad \Delta(k, \ell)$ s.t. $|\mathcal{U}_k| < N_{u_k}$
10: Allocate ℓth ant. to the kth user: $\mathcal{U}_k = \mathcal{U}_k \cup \{\ell\}$;
11: Remove ℓth ant. from remained ant. and set: $\mathcal{N} = \mathcal{N} - \{\ell\}$;
12: Update $\mathbf{H}_k = \left[\mathbf{H}_{\mathcal{U}_k}, \mathbf{H}_{\mathcal{C}}\right]$;
13: Update kth row of Δ from steps 2 to 7;
14:**Until** $|\mathcal{N}| = 0$
Output: $\mathcal{U}_k, \qquad \forall k = 1, \ldots, K$;

TABLE 1. Implementation Details of the DNN.

Layer Name	Output Dim.	Activation Func.
Input Layer	$K \times N_t \times 3$	–
Dense Layer 1	$KN_t \times 1$	ReLU
Dense Layer 2	$KN_t \times 1$	ReLU
Output Layer	$K \times N_t$	Sigmoid

IV. PROPOSED DNN FOR THE DYNAMIC PFC ARCHITECTURE

In this section, we present the use of a deep neural network trained to predict the results of the algorithms 1 and 2 directly from the CSI. The configuration of the neural network we used is illustrated in Table 1; it has four layers including the input layer, two dense layers, and the output layer. The hidden layers deploy KN_t units and the rectified linear unit (ReLU) function, as the activation function, but the output layer uses the sigmoid function.

The input layer is fed by a 3D matrix denoted as $\mathbf{X} \in \mathbb{R}^{K\times N_t\times 3}$, with $(i, j, :)$-th entry, a vector of size 3 containing

absolute, real, and imaginary parts of the (i, j)-th entry of the channel matrix. The training label matrix $\mathbf{Y}$, of size $K \times N_t$, shows the connection between RF-chains and antennas. $\mathbf{Y}$ is constructed in the simulation using the previously proposed algorithms as:

$$\mathbf{Y}(k,n) = \begin{cases} 1 & n \in \mathcal{U}_k \quad \text{or} \quad n \in \mathcal{C} \\ 0 & otherwise \end{cases} \tag{29}$$

In the training phase, the DNN is trained end-to-end in a supervised manner. More specifically, a dataset of the channel matrix and the corresponding connection matrix $\mathbf{Y}$ are used to train the DNN to predict the connections for a given input channel matrix.

A. DATASET

We have generated L_p realizations of the channel matrix with different user locations. For each of which, L_n noisy channel matrices are generated as:

$$\mathbf{H}^{(l_p,l_n)} = \mathbf{H}^{(l_p)} + \mathbf{z}, \tag{30}$$

where $\mathbf{z}$ is a complex Gaussian noise matrix, whose entries are i.i.d. and follow the distribution $\mathcal{CN}(0, \sigma_z^2)$. For each noiseless channel matrix $\mathbf{H}^{(l_p)}$, the sets $\mathcal{C}$ and $\mathcal{U}_k$ are constructed, thanks to Algorithm 1 and Algorithm 2, respectively. Then the training label, $\mathbf{Y}^{(l_p)}$, is calculated from (29). Furthermore, the $\mathbf{Y}^{(l_p)}$ is considered as the training label for all noisy forms of the $\mathbf{H}^{(l_p)}$, which gives the input-output pairs of the training data.

B. TRAINING PHASE

During the training process, the DNN is fed by the training data generated for $L_p = 1000$ channel realizations. To account for different channel characteristics, for each channel realization, $L_n = 99$ noisy channels are generated by adding synthetic noises for different powers of $\text{SNR}_{\text{TRAIN}} \in \{15, 20, 25\}$ (33 noisy channels of each power), where $\text{SNR}_{\text{TRAIN}} = 20\log_{10}\left(\frac{|H^{(l_p)}(i,j)|^2}{\sigma_z^2}\right)$. In the training process, 70% of all generated data is selected as the training set and the remaining as the validation set.

C. POST-PROCESSING PHASE

The trained DNN accepts the input of size $K \times N_t \times 3$, and predicts the matrix $\widehat{\mathbf{Y}}$, which contains the connection "probabilities". To satisfy the connectivity constraints, post-processing is required to obtain the sets $\mathcal{C}$ and $\mathcal{U}_k$. We suggest a straightforward solution, where the N_c columns of $\widehat{\mathbf{Y}}$ with the largest sum-value are assigned to set $\mathcal{C}$. In fact, it consists of selecting the antennas with the most connectivity to all users. Next, to assign the antennas to the users, a max-min simple approach is applied to add the number ℓ to the set $\mathcal{U}_k$. Intuitively, we first look for the less significant antenna, and we connect it to the user with the most probability of connection. For that, we sum up all the columns of $\widehat{\mathbf{Y}}$ and select the minimum. Then, in that column, we select the user with the most probability of connection. The post-processing is summarized in Algorithm 3.

Algorithm 3 Post-Processing for the Proposed DNN

Input: $\widehat{\mathbf{Y}}$;
$\mathcal{N} = [N_t]$;
$\mathcal{U}_k = \{\}, \qquad \forall k = 1, \ldots, K$;
$\mathcal{C} = \{\}$;
Repeat
  $\ell = \text{argmax}_{\ell \in \mathcal{N}} \quad \sum_{k=1}^{K} \widehat{\mathbf{Y}}(k, \ell)$
  $\mathcal{C} = \mathcal{C} \cup \{\ell\}$;
  $\mathcal{N} = \mathcal{N} - \{\ell\}$;
Until $|\mathcal{C}| = N_c$
Repeat
  $\ell = \text{argmin}_{\ell \in \mathcal{N}} \sum_{k=1}^{K} \widehat{\mathbf{Y}}(k, \ell)$
  $k = \text{argmax}_k \quad \widehat{\mathbf{Y}}(k, \ell)$
    s.t. $\quad |\mathcal{U}_k| < N_{u_k}$;
  $\mathcal{U}_k = \mathcal{U}_k \cup \{\ell\}$;
  $\mathcal{N} = \mathcal{N} - \{\ell\}$;
Until $|\mathcal{N}| = 0$
Outputs: $\mathcal{C}$, and $\mathcal{U}_k, \qquad \forall k = 1, \ldots, K$;

TABLE 2. Computation Times (in Milliseconds).

No. FACs, N_c	10	25	50	100	113
Alg. 1 & 2	1317.9	1219.3	1183.2	822.4	792.2
Proposed DNN	8.5	8.3	8.1	7.4	7.3

The complexity of the proposed algorithms is predominantly dependent on the computational complexity of the matrix inversion and sorting data, which are of the order $O(n^3)$ and $O(n \log n)$, respectively. In Algorithm 1, the computational complexity grows linearly by the number of iterations, N_c. It calculates $N_t - i$ matrix inversions in the ith iteration, so the total number of computing matrix inversion is $N_c\,(N_t - (N_c - 1)/2)$. Furthermore, in each iteration, it sorts the trace values in step 3, where the maximum number of values is N_t. Therefore, the computational complexity of Algorithm 1 is $\mathcal{O}\left(N_c\left((N_t - (N_c - 1)/2)\,K^3 + N_t \log N_t\right)\right)$. Algorithm 2 includes two loops; the complexity of the first loop depends on calculating N_U matrix inversions; the complexity of the second loop grows linearly by N_U, and calculating $N_U^2/2$ matrix inversion. Furthermore, it sorts values of matrix Δ in each iteration. Therefore, the computational complexity of Algorithm 2 is $\mathcal{O}\left(N_U\left(K^4 + N_U/2K^3 + N_U K \log\left(N_U K\right)\right)\right)$. Algorithm 3 sorts a vector of the size N_t in each iteration, therefore its complexity is of the order $\mathcal{O}\left(N_t^2 \log N_t\right)$.

Furthermore, the computation time of the algorithms is presented in Table 2 for a different number of FACs when the BS equipped with 256 antennas to serve 10 users. It shows the proposed DNN significantly reduces the computation time.

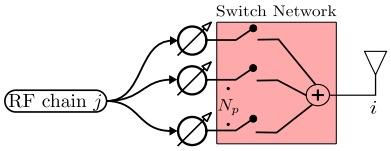

FIGURE 4. Signal flow from the RF-chain *j* to the antenna *i*.

V. SIMULATION RESULTS

In this section, the simulation results are presented to show the relation between performance and hardware complexity in a multi-user massive MISO system, and to compare them with the FC and GC strategies [21]. In the following simulations, we consider Saleh-Valenzuela channel, which models the propagation environment as a geometric channel with N_{cl} paths [31]. The channel vector between user k and the BS, $\mathbf{h}_k$, is expressed as:

$$\mathbf{h}_k = \sqrt{\frac{N_t}{N_{cl}}} \sum_{i=1}^{N_{cl}} \alpha_i \mathbf{a}_t \left(\phi_i, \theta_i\right)^H, \tag{31}$$

where α_i represents the channel gain of the ith path. We assume that all the paths have the same average power, $\alpha_i \sim \mathcal{CN}(0, 1)$. The ϕ_i and θ_i are angles of departure uniformly distributed across 60 degrees in the azimuth domain and 20 degrees in elevation [15]. The $\mathbf{a}(\phi_i, \theta_i)$ represents the array response vector of transmitter antenna array. For a N by M uniform planar array (UPA), $\mathbf{a}(\phi_i, \theta_i)$ is given by [32]

$$\mathbf{a}(\phi_i, \theta_i) = \frac{1}{\sqrt{NM}} \Big[1, \ldots, e^{J\frac{2\pi}{\lambda_c} d(n \sin(\phi_i)\sin(\theta_i) + m\cos(\theta_i))}, \ldots, e^{J\frac{2\pi}{\lambda_c} d((N-1)\sin(\phi_i)\sin(\theta_i) + (M-1)\cos(\theta_i))}\Big]^{\mathrm{T}}, \tag{32}$$

where λ_c is the the wavelength, d is the inter-element spacing, and $0 \le n < N$ and $0 \le m < M$ are the y and z indexes of antenna elements, respectively.

implementation model:To simplify the hardware complexity, the architecture of the analog beamforming is presented in [21] and [23]. In this architecture, the complex weightings are approximated by N_p fixed phase shifters (FPS) and a switch network presented in Figure 4. The advantage is that the generated phases are shared among all the other coefficients in the same row of the matrix $\mathbf{F}$. It is shown in [23] that with only 11 FPSs per RF-chain, a quasi-optimal performance can be obtained. In our simulations, the transmitter is equipped with a 16×16 UPA with half-wavelength spacing between elements, $N_p = 11$, and $N_{RF} = K$. Furthermore, we assume that there is no priority between users, so the same number of antennas are assigned to them.

A. SPECTRAL EFFICIENCY

To analyze the performance of the proposed architecture, the sum-rate criterion is considered:

$$R_s = \sum_{k=1}^{K} \log_2 \left(1 + \frac{\rho \left|\mathbf{h}_k \mathbf{f}_k\right|^2}{\rho \sum_{k'=1, k' \neq k}^{K} \left|\mathbf{h}_k \mathbf{f}_{k'}\right|^2 + 1}\right). \tag{33}$$

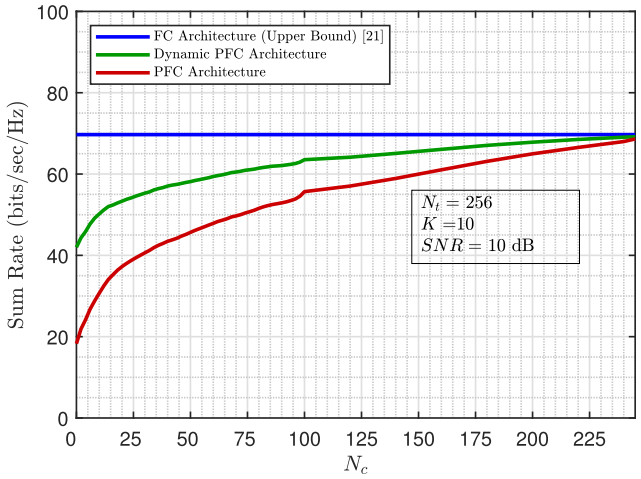

FIGURE 5. The sum-rate as a function of N_c for $N_t = 256$, $K = 10$ and $SNR = 10$ dB.

Figure 5 illustrates the sum-rate as a function of N_c, going from PC to FC strategies. As it can be seen, for small values of N_c, the performance sharply increases by assigning more antennas to the set of FCA's, resulting in more complex hardware. The designer can select the desired trade-off between complexity and performance. As expected, the dynamic PFC significantly improves the SE for small values of N_c; the price to pay is the added switch matrix before the antennas.

B. ENERGY EFFICIENCY

To put the energy consumption into perspective, we define the EE as $\xi \triangleq \frac{R_s}{P_{tot}}$, where P_{tot} is the total power consumption introduced as:

$$P_{tot} = P + P_{BB} + N_{RF} P_{RF} + N_{RF} N_p P_{PS}^a + N_{nz} N_p P_{sw} \tag{34}$$

In this equation, P is the transmit power; P_{PS}^a, P_{BB}, and P_{RF} are, respectively, the powers consumed by a PS, by the baseband processor, and by an RF-chain; P_{sw} represents the power consumed by a switch. In this simulation, their values are set to: $P_{PS}^F = 10mW$, $P_{PS}^Q = 30mW$, $P_{PS}^C = 50mW$, $P_{BB} = 200mW$, $P_{RF} = 300mW$, and $P_{sw} = 5mW$ [22]. Also, the number of switch networks N_{nz} in the FC, GC, PFC, and dynamic PFC architectures are KN_t, KN_t/g, $N_U + KN_c$, and $2N_t + N_c(K-1)$, respectively.

Figure 6 illustrates the energy efficiency as a function of the number of FCAs. This shows that the maximum energy efficiency is achieved by considering no antenna as FCA in dynamic PFC architecture, while the fixed PFC architecture requires a few FCAs.

C. DNN BASED ANTENNA ASSIGNMENT

In Figure 7 and Figure 8, the trained DNN is used for antennas assignment in the dynamic PFC architecture. To have a fair comparison with the GC architecture [18], we consider approximately the same hardware complexity for both

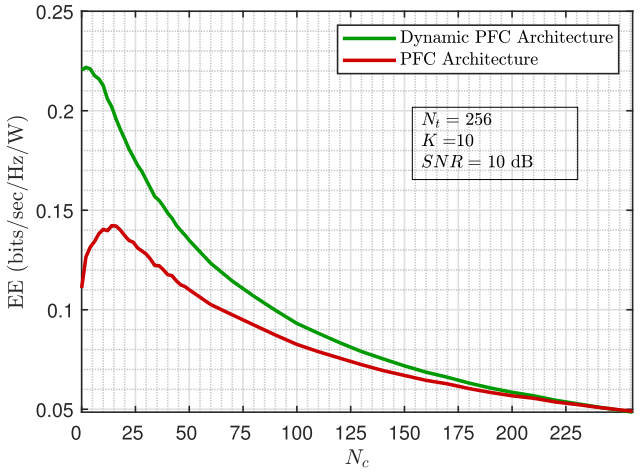

FIGURE 6. Energy efficiency achieved by different values of N_c, when $K = 10$, and $SNR = 10$ dB.

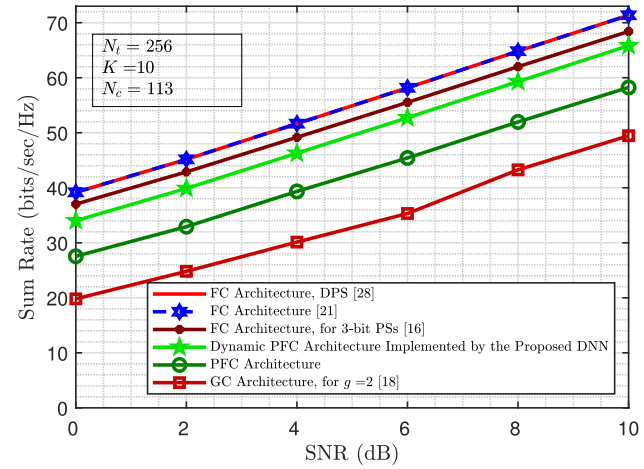

FIGURE 7. Sum-rate versus SNR, with $K = 10$, $N_c = 113$ in the PFC, $g = 2$ in the GC.

structures. Therefore, the number of shared antennas is determined by:

$$N_c = \lfloor \frac{N_t(K-g)}{g(K-1)} \rfloor, \tag{35}$$

where $\lfloor . \rfloor$ denotes the floor function. Since the performance of the GC strategy decreases by increasing g, to achieve comparable performance, the number of groups is set to the minimum, i.e., $g = 2$. Figure 7 reveals a non-negligible gain in the sum-rate for the dynamic PFC compared to the GC architecture. The performance of FC architecture deploying double PS (DPS) and quantized PS is also illustrated. For instance, at 10 dB of SNR, the PFC and dynamic PFC architectures obtain 81.62% and 92.26% of the FC rate, respectively, while the GC strategy achieves 69% of that. Figure 8 presents the energy efficiency versus SNR. It indicates that the energy efficiency of the proposed architecture is significantly better than that of the FC and the GC structures. It reveals that even though the FC structure has better spectral efficiency, it has

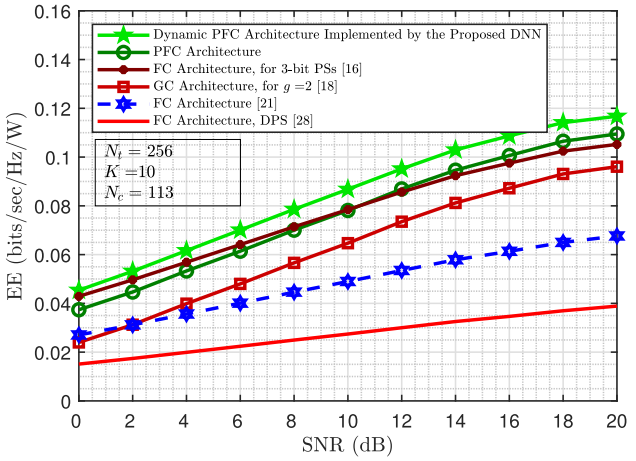

FIGURE 8. Energy efficiency achieved by different values of SNR, when $K = 10$, $N_c = 113$ in the PFC, and $g = 2$ in the GC.

a lower energy efficiency because of the number of deployed switches.

VI. CONCLUSION

The present paper considered the downlink transmission in multi-user massive MISO systems, where the analog BF is performed at the BS. Regarding the impacts of RF-paths on performance and hardware complexity, we proposed a cost-efficient architecture with $N_t + 1$ distinct levels of complexity/performance trade-off. Particularly, the antenna elements are divided into two groups. N_c antennas are connected to all the RF-chains through the FC architecture, and each of the rest is connected to only one RF-chain, as the PC strategy, reducing hardware complexity. The lack of connection between all the RF-chain/antenna pairs imposes constraints on the form of the precoder matrix, i.e., zero elements, which makes the optimization procedure more complicated. We proposed an analytic solution for the optimization of the precoding matrix by using the zero-forcing approach. To improve the performance, the dynamic antenna assignment is presented, which poses a complex combinatorial optimization problem. A suboptimal greedy solution is given for antenna selection with reasonable complexity. To further simplify the implementation for real-time applications, we proposed a machine learning approach that gets the CSI and gives directly the antenna assignment matrix, thanks to a post-processing simple algorithm. Finally, we presented promising simulation results on the performance of the proposed architecture for the mmWave channels.

REFERENCES

[1] E. Dahlman, S. Parkvall, and J. Skold, *5G NR: The Next Generation Wireless Access Technology*. New York, NY, USA: Academic, 2020.
[2] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1617–1655, 3rd Quart., 2016.
[3] A. Ghosh, R. Ratasuk, B. Mondal, N. Mangalvedhe, and T. Thomas, "LTE-advanced: Next-generation wireless broadband technology," *IEEE Wireless Commun.*, vol. 17, no. 3, pp. 10–22, Jun. 2010.

[4] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 101–107, Jun. 2011.
[5] A. Ghosh, A. Maeder, M. Baker, and D. Chandramouli, "5G evolution: A view on 5G cellular technology beyond 3GPP release 15," *IEEE Access*, vol. 7, pp. 127639–127651, 2019.
[6] H. Holma, A. Toskala, and T. Nakamura, *5G Technology: 3GPP New Radio*. Hoboken, NJ, USA: Wiley, 2020.
[7] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.
[8] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.
[9] S. A. Busari, K. M. S. Huq, S. Mumtaz, L. Dai, and J. Rodriguez, "Millimeter-wave massive MIMO communication for future wireless systems: A survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 2, pp. 836–869, 2nd Quart., 2018.
[10] R. W. Heath, N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 436–453, Feb. 2016.
[11] W. Roh, J.-Y. Seol, J. Park, B. Lee, J. Lee, Y. Kim, J. Cho, K. Cheun, and F. Aryanfar, "Millimeter-wave beamforming as an enabling technology for 5G cellular communications: Theoretical feasibility and prototype results," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 106–113, Feb. 2014.
[12] X. Yang, M. Matthaiou, J. Yang, C.-K. Wen, F. Gao, and S. Jin, "Hardware-constrained millimeter-wave systems for 5G: Challenges, opportunities, and solutions," *IEEE Commun. Mag.*, vol. 57, no. 1, pp. 44–50, Jan. 2019.
[13] I. Ahmed, H. Khammari, A. Shahid, A. Musa, K. S. Kim, E. De Poorter, and I. Moerman, "A survey on hybrid beamforming techniques in 5G: Architecture and system model perspectives," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 3060–3097, 4th Quart., 2018.
[14] A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, S. L. H. Nguyen, L. Li, and K. Haneda, "Hybrid beamforming for massive MIMO: A survey," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 134–141, Sep. 2017.
[15] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath Jr., "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.
[16] 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.
[17] X. Yu, J. 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, Feb. 2016.
[18] X. Yu, J. Zhang, and K. B. Letaief, "A hardware-efficient analog network structure for hybrid precoding in millimeter wave systems," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 2, pp. 282–297, May 2018.
[19] A. Alkhateeb and R. W. Heath Jr., "Frequency selective hybrid precoding for limited feedback millimeter wave systems," *IEEE Trans. Commun.*, vol. 64, no. 5, pp. 1801–1818, May 2016.
[20] O. El Ayach, R. W. Heath Jr., S. Rajagopal, and Z. Pi, "Multimode precoding in millimeter wave MIMO transmitters with multiple antenna sub-arrays," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2013, pp. 3476–3480.
[21] J. Beiranvand, V. Meghdadi, C. Menudier, and J. P. Cances, "An efficient low-complexity method to calculate hybrid beamforming matrices for mmWave massive MIMO systems," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 1239–1248, 2021.
[22] H. Li, M. Li, and Q. Liu, "Hybrid beamforming with dynamic subarrays and low-resolution PSs for mmWave MU-MISO systems," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 602–614, Jan. 2020.
[23] J. Beiranvand, V. Meghdadi, C. Menudier, and J.-P. Cances, "How many fixed phase shifters are needed in a hybrid BF structure?" in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2022, pp. 444–449.
[24] N. Song, T. Yang, and H. Sun, "Overlapped subarray based hybrid beamforming for millimeter wave multiuser massive MIMO," *IEEE Signal Process. Lett.*, vol. 24, no. 5, pp. 550–554, May 2017.
[25] S. Park, A. Alkhateeb, and R. W. Heath Jr., "Dynamic subarrays for hybrid precoding in wideband mmWave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 2907–2920, May 2017.
[26] L. Yan, C. Han, and J. Yuan, "A dynamic array-of-subarrays architecture and hybrid precoding algorithms for terahertz wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 9, pp. 2041–2056, Sep. 2020.
[27] J.-C. Guo, Q.-Y. Yu, W.-X. Meng, and W. Xiang, "Energy-efficient hybrid precoder with adaptive overlapped subarrays for large-array mmWave systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1484–1502, Mar. 2020.
[28] X. Yu, J. Zhang, and K. B. Letaief, "Doubling phase shifters for efficient hybrid precoder design in millimeter-wave communication systems," 2019, *arXiv:1905.10624*.
[29] H. Avron and C. Boutsidis, "Faster subset selection for matrices and applications," *SIAM J. Matrix Anal. Appl.*, vol. 34, no. 4, pp. 1464–1499, Jan. 2013.
[30] F. De Hoog and R. Mattheij, "Subset selection for matrices," *Linear Algebra Appl.*, vol. 422, nos. 2–3, pp. 349–359, 2007.
[31] T. S. Rappaport, R. W. Heath Jr., R. C. Daniels, and J. N. Murdock, *Millimeter Wave Wireless Communications*. London, U.K.: Pearson Education, 2015.
[32] C. A. Balanis, *Antenna Theory: Design and Analysis*, 3rd ed. Hoboken, NJ, USA: Wiley, 1997, pp. 978–979.

JAMAL BEIRANVAND received the M.S. degree from Semnan University, Iran, in 2018. He is currently pursuing the Ph.D. degree in telecommunication engineering with the XLIM Research Laboratory, University of Limoges, France. His research interests include wireless communication systems design, massive MIMO systems, millimeter wave communications, mathematical optimization, fundamental mathematics, information theory, and audio and speech processing.

VAHID MEGHDADI received the B.Sc. and M.Sc. degrees from the Sharif University of Technology, Tehran, Iran, in 1988 and 1991, respectively, and the Ph.D. degree from the University of Limoges, France, in 1998. He has been a Professor with the Department of Electronic and Telecommunication, ENSIL, University of Limoges, since 2000, and a Researcher with the CNRS XLIM Laboratory, Limoges, France. He served as the Scientific Manager for more than ten research projects in the field of information and communications technology (ICT). His research interests include telecommunication systems, coding, network coding, cooperative communications, sensor networks, and massive MIMO systems.

CYRILLE MENUDIER (Member, IEEE) was born in France, in 1981. He received the M.Sc. degree in high-frequency telecommunications from the University of Limoges, the engineer degree from in electronics from ENSIL, in 2004, and the Ph.D. degree in telecommunications from the XLIM Research Laboratory, University of Limoges, in 2007. He then held a postdoctoral position at CNES (French Space Agency), Toulouse, until 2009, where he worked on reconfigurable reflectarray antennas. He is currently an Associate Professor (HDR) of RF systems axis with the XLIM Research Laboratory. His research interests include reconfigurable antennas, active antennas and phased arrays, reflectarrays, parasitic element antennas, and mutual coupling effects.

JEAN PIERRE CANCES (Member, IEEE) received the graduate degree from the Ecole Nationale Supérieure des Télécommunications, in 1990, and the aggregation teaching degree, in 1993. Since 2006, he has been a Full Professor of digital signal processing for communication systems with ENSIL, University of Limoges. His current research interests include wireless sensor systems (WSNs), NOMA multi-user detection, and resource allocation in 5G and beyond.