IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 20, 2019, accepted August 4, 2019, date of publication August 22, 2019, date of current version September 10, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2936911

Robust Beamforming and User Clustering for Guaranteed Fairness in Downlink NOMA With Partial Feedback

MOHANAD M. AL-WANI[1], ADUWATI SALI[1], NOR K. NOORDIN[1], SHAIFUL J. HASHIM[1], CHEE YEN LEOW[2], AND IOANNIS KRIKIDIS[3], (Fellow, IEEE)

[1]Wireless and Photonic Networks Research Centre of Excellence, Department of Computer and Communication Systems Engineering, Faculty of Engineering, Universiti Putra Malaysia, Serdang 43400, Malaysia
[2]Wireless Communication Centre, School of Electrical Engineering, Universiti Teknologi Malaysia, Johor 81310, Malaysia
[3]Electrical and Computer Engineering Department, University of Cyprus, 1678 Nicosia, Cyprus

Corresponding author: Aduwati Sali (aduwati@upm.edu.my)

This work was supported by "ATOM"-Advancing the State of the Art of MIMO (Proj. No.: 690750-ATOM-H2020-MSCA-RISE-2015, UPM: 6388800-10801), in part by: NOMA-MIMO: Optimizing 5G Wireless Communication Performance Based on Hybrid NOMA with Partial Feedback for Multiuser MIMO (GP-IPS/2018/9663000, Vote No: 9663000), in part by: EMOSEN-Energy Efficient MIMO-Based Wireless Transmission for SWIPT-Enabled Network (Vote No: 9671600) (UPM/800-3/3/1/GPB/2019/9671600), and co-funded in part by the European Regional Development Fund and the Republic of Cyprus through the Research Promotion Foundation, under the project INFRASTRUCTURES/1216/0017.

ABSTRACT In this paper, a downlink multiuser non-orthogonal multiple access (NOMA) with full and partial channel state information (CSI) feedback is considered. We investigate beam design and user clustering from the throughput-fairness trade-off perspective. To enhance this trade-off, two proportional fairness (PF) based scheduling algorithms are proposed, each has two stages. The first algorithm is based on integrating the maximum product of effective channel gains and the maximum signal to interference ratio with the PF principle (PF-MPECG-SIR), to select the strong users in the first stage and the weak users in the second stage. This algorithm is designed to maximize the throughput with moderate fairness enhancement. Whereas, in the second algorithm, the MPECG and the maximum correlation are combined within the PF selection criterion (PF-MPECG-CORR) in order to maximize the fairness with a slight degradation in the total throughput. In addition, we present an optimal power allocation that can achieve a high data rate for the overall system without sacrificing the sum-rate of weak users under full and partial CSI. Simulation results show that the proposed PF-MPECG-CORR can significantly improve the fairness up to 50.82% and 44.90% with only 0.42% and 1.13% degradation in the total throughput, for full and partial CSI, respectively. All these performance gains are achieved without increasing the computational complexity.

INDEX TERMS 5G, MPECG, NOMA, partial channel state information, proportional fairness, power allocation, zero-forcing beamforming.

I. INTRODUCTION

In recent years, non-orthogonal multiple access (NOMA) has been considered as a pivotal technique for the upcoming fifth-generation (5G) wireless networks, due to many reasons such as: high spectral efficiency, massive connectivity by taking advantages of user grouping and multiplexing in the same time/frequency resources [1], [2]. The basic principle of NOMA is to superpose multiple users signals in the power domain (superposition coding) at the transmitter side and performing successive interference cancellation (SIC) at the receiver of the strong or the near user to remove inter-user interference from the desired signal [3], [4].

In the earlier works on NOMA, different types of resource allocation schemes have been studied, the sum-rate maximization was the most commonly adopted objective, and there are several related works [5]–[7].

In [8], an optimal scheme user pairing and power allocation (PA) scheme were proposed for the NOMA system with the proportional fairness objective.

The associate editor coordinating the review of this article and approving it for publication was Qilian Liang.

In [9], proportional fairness scheduling (PFS) was derived for NOMA with two users under two different criteria: the maximum of the sum rate and the maximum of the minimum rate to make PF with a small variation of transmission rates.

In [10], a low-complexity waterfilling-based power allocation technique, incorporated within the PF scheduler is proposed to maximize the achieved average throughput and guaranteeing a high level of fairness for NOMA system.

In downlink multi-user systems, beamforming is an important and essential technique that can be used to improve system performance in terms of throughput and the number of served users. NOMA was first integrated with zero-forcing beamforming (NOMA-ZFBF) in [11] by B. Kim to enhance the sum capacity. By selecting two correlated users with a maximum channel gain difference to form a 'NOMA cluster'. Each NOMA cluster can be served with a single beam by applying the principle of NOMA between the users of that cluster. However, their results showed that the weak users' sum-rate is decreased with increasing the number of candidate users.

In [12], powerful user selections algorithms were proposed. These algorithms significantly improved the weak users' and total system sum-rates of NOMA-ZFBF compared to that in [11]. Strong users are selected using semiorthogonal user selection (SUS) algorithm [13], which ensures the smallest possible inter-cluster interference (ICI). Weak users are then chosen using the designed beams from the strong users' channels, and the selection criterion is based on the maximum achieved signal to interference ratio (SIR). This algorithm is referred here as the ''SUS-SIR'' algorithm. In [14], to improve the throughput-fairness trade-off for downlink NOMA-ZFBF, a two-stage user scheduling algorithm was proposed based on PF. This algorithm is referred here as the ''PF-SUS-SIR'' algorithm.

In [15], beam design and-user-scheduling based on the Pareto-optimality algorithm was proposed to control the rates of strong and weak users in downlink multiuser NOMA. However, the performance of SUS-SIR algorithm in [12] still gives a higher throughput for the total system and weak users.

Most of the existing works on NOMA are based on full (perfect) CSI assumption at the transmitter side. In practice, this assumption is difficult to obtain, due to many limitations like high mobility of users, channel estimation errors, and feedback delays. Therefore, it is important to investigate whether the advantages of NOMA still appears under partial (limited) CSI condition and to consider a robust resource allocation schemes under partial CSI conditions.

The authors in [16] proposed a dynamic user scheduling and derive a closed-forms approximation of outage probability and the net throughput based on leveraging limited feedback.

In [17], various beamforming techniques were developed for downlink NOMA system with full CSI and norm-bounded channel uncertainties to meet the required quality of service (QoS) for all users. However, user scheduling was not considered in the system, which yields more inter-user interference as the number of served users increases and hence, the capacity of this system decreases as the number of served users increases.

In [18], zero-forcing and random beamforming techniques were investigated in downlink multiuser NOMA system with limited CSI. In addition, user selection and power allocation schemes were proposed to improve the total sum-rate of NOMA system.

The contributions of this paper are summarized as follows:

1) We propose two powerful and fair user clustering algorithms for NOMA-ZFBF system based on PF with two types of resource allocation: the first algorithm combines the maximum product of effective channel gains (MPECG) [19], and the maximum signal to interference ratio within the PF selection criterion (PF-MPECG-SIR) which is able to maintain the total system capacity (achieved by SUS-SIR [12]) with a moderate enhancement in user fairness. The second clustering algorithm integrates the MPECG and the maximum correlation into the PF selection criterion (PF-MPECG-CORR), and it's designed to maximize fairness with a slight degradation in the total throughput compared to the PF-MPECG-SIR algorithm.
2) In addition to full CSI, we test our proposed algorithms with a partial CSI scenario with a different numbers of feedback bits per user B. The quantized channel (partial feedback channel) is modeled based on the random vector quantization (RVQ) quantization technique, which considered an asymptotic optimal quantization technique. To the best of our knowledge, there are few works on NOMA-ZFBF system have used the RVQ approach such as in [16] and [18]. We also provide the optimal number of feedback bits B needed for each user to reach the full CSI capacity of the NOMA system using RVQ codebook, which has not been considered in the previous works.
3) Enhance fairness between users with maintaining the total system throughput. In other words, achieve a better throughput-fairness tradeoff. This enhancement is achieved without increasing the system complexity.
4) We propose a new power allocation strategy that ensures the weak users' sum capacity under full and partial CSI. Unlike other techniques, such as [18], which focused on maximizing the total system throughput and sacrificing in weak users' sum capacity, our proposed power allocation can guarantee a balance between maximizing the total throughput and maintaining a high data rate for the weak users under both full and partial CSI.

The rest of this paper is organized as follows: the system model is described in Section II. Section III presents the received signal model for NOMA-ZFBF users. The CQI feedback model is presented in Section IV and the proportional fairness scheduling is discussed in Section V. In Section VI

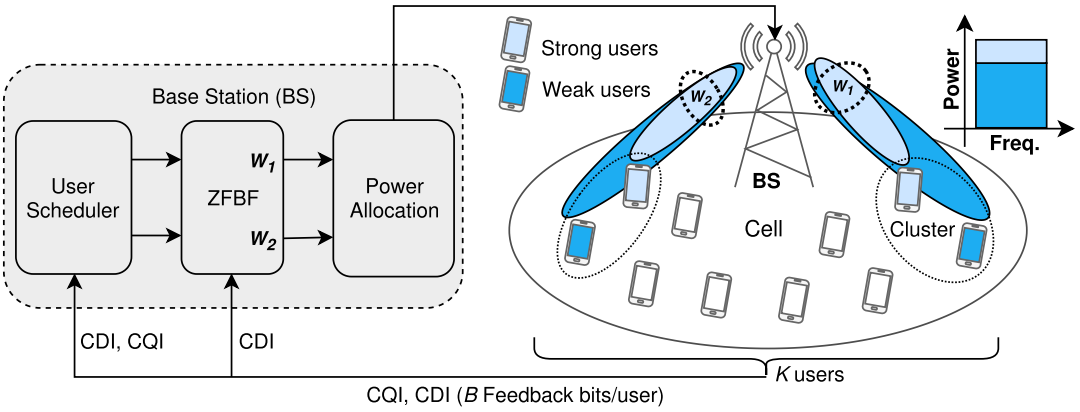

FIGURE 1. System model of downlink NOMA with beamforming.

the transmit power allocation is disused. Section VII presents the proposed clustering algorithms. The computational complexity analysis is given in Section VII. Simulation results is presented in Section IX. Finally, the conclusions are discussed in section X.

A. NOTATION

We use lowercase boldface letters for vectors and uppercase boldface letters for matrices. $\mathbb{E}(\cdot)$ correspond to the statistical expectation for random variables, $(\cdot)^T$, $(\cdot)^H$ and $\|\cdot\|$ denote the transpose, conjugate transpose operations, and the Frobenius norm operation, respectively.

II. SYSTEM MODEL

We consider the downlink multi-user BF-NOMA system in a single-cell network, as shown in Fig. 1. The BS is equipped with N_t antennas and communicates with K single-antenna users. The candidate users set is defined as K and ($|U| = K$). In NOMA-ZFBF system, the BS can simultaneously serve N_t clusters with two or more users per cluster served by a single beam. We assume that each cluster contains two users as [11], [12]. NOMA principle is applied between the clustered users, which have different channel gains, the user with a high channel gain is known as the strong user (denoted with "k_1''), and the other, with small channel gain is known as the weak user (denoted with "k_2''). At each time slot t, the BS transmits the superimposed symbols in power domain to the strong and weak users in each cluster. The transmitted superimposed symbols to the kth cluster x_k can be written as

$$x_k = \sqrt{\alpha_k} s_{k1} + \sqrt{1-\alpha_k} s_{k2}, \tag{1}$$

where α_k and $(1-\alpha_k)$ are the PA coefficients for the strong and weak user, respectively, such that $(1-\alpha_k) \geq \alpha_k$ and $\alpha_k + (1-\alpha_k) = 1$, s_{k1} and s_{k2} are the transmitted symbols to the strong and weak user, respectively. Therefore, the received signal y_{kn} at the two users in the kth cluster is given by

$$y_{kn} = \sqrt{P}\boldsymbol{h}_{kn}\boldsymbol{w}_k x_k + \sqrt{P}\sum_{\substack{j=1 \\ j\neq k}}^{N_t} \boldsymbol{h}_{kn}\boldsymbol{w}_j x_j + n_{kn}, \quad \text{for } n = 1, 2 \tag{2}$$

where $\boldsymbol{h}_{kn} \in \mathbb{C}^{1\times N_t}$ denotes the channel vector which is modeled as a flat Rayleigh fading channel with zero mean and unit variance, $\boldsymbol{w}_k \in \mathbb{C}^{N_t\times 1}$ is the ZFBF vector for the cluster k which is designed based on the imperfect feedback channels of the strong users and n_{kn} is complex additive white Gaussian noise (AWGN) with distribution $\mathcal{CN}(0,\ \sigma_n^2)$.

A. PARTIAL FEEDBACK MODEL

As shown in Fig. 1, each mobile user quantizes its normalized channel direction information (CDI) into B bits and feeds back them to the BS. The CSI quantization is performed using the well-recognized random vector quantization (RVQ) scheme, in which the precoding matrix is selected from a random codebook consists of 2^B quantization vectors i.e. $Q \triangleq \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{2^B}\}$. These vectors are randomly and independently generated from the isotropic distribution of N_t-dimensional unit sphere [18], [20]. The quantized CDI channel (partial feedback channel) vector $\hat{\boldsymbol{h}}_k$ is chosen according to the following decision rule:

$$\hat{\boldsymbol{h}}_k = \arg\min_{j\in Q} \sin^2\left(\angle(\tilde{\boldsymbol{h}}_k, \boldsymbol{w}_j)\right), \tag{3}$$

Note that if the CDI is full (i.e., $\hat{\boldsymbol{h}}_k = \tilde{\boldsymbol{h}}_k = \boldsymbol{h}_k/\|\boldsymbol{h}_k\|$), the multiuser interference would be totally removed by ZFBF. Under quantized CDI, however, mutual interference among users cannot completely eliminated because the beamforming vectors are designed orthogonal to the quantized channel $\hat{\boldsymbol{h}}_k$ and not the actual channel $\boldsymbol{h}_k$.

To realize the effect of partial feedback, let us firstly review some basic results for a single user multiple-input-single-output (MISO) system, In case of full CSI, the transmitter can optimally beamform along the channel vector $\boldsymbol{h}$, and

therefore, the corresponding rate is [20], [21]

$$R_{CSI} = \log(1 + P\|\boldsymbol{h}\|^2), \tag{4}$$

Under quantized CDI feedback using RVQ quantization, the average rate of this system becomes

$$\begin{aligned} R_{CDI} &= \mathbb{E}\left[\log(1 + P\|\boldsymbol{h}\|^2 \cos^2\theta\right] \\ &= \mathbb{E}\left[\log(1 + P\|\boldsymbol{h}\|^2(1 - \sin^2\theta\right)] \\ &\approx \mathbb{E}\left[\log\left(1 + P\|\boldsymbol{h}\|^2(1 - 2^{-\frac{B}{N_t-1}})\right)\right], \end{aligned} \tag{5}$$

where $\theta = \angle(\boldsymbol{h}, \hat{\boldsymbol{h}})$ is the angle between the $\boldsymbol{h}$ and $\hat{\boldsymbol{h}}$, and $\sin^2\theta$ is the quantization error which can be upper bounded by $2^{-\frac{B}{N_t-1}}$ in the last approximation. Therefore, partial feedback results in sum-rate degradation of approximately $10\log_{10}(1 - 2^{-\frac{B}{N_t-1}})$ decibels relative to full CSI [21].

B. ZERO-FORCING BEAMFORMING

ZFBF is a downlink precoding technique which is widely used in wireless systems to achieve high throughput by simultaneously transmitting multiple data streams for different users. In NOMA-ZFBF system, ZF beams are designed based on the strong users' channels $\boldsymbol{h}_{k1}$. These beams are then used to precode the transmitted superimposed symbols for the scheduled strong and weak users in each cluster. In case of full CSI, each user can completely eliminate the interference comes from other NOMA clusters (referred to as inter-cluster interference) by multiplying the user channel with the corresponding BF vector such as $\boldsymbol{h}_{k1}\boldsymbol{w}_j = 0$ for $k \neq j$. However, because we consider partial feedback system, the BF matrix is actually designed based on the quantized CDIs of the strong users and not the actual channels. Therefore, each BF vector $\boldsymbol{w}_k$ is chosen to satisfy:

$$\hat{\boldsymbol{h}}_{k1}\boldsymbol{w}_j = 0, \forall j \neq k, \tag{6}$$

Let $\hat{\boldsymbol{H}}(S_{su}) = [\hat{\boldsymbol{h}}_{11}^T, \cdots, \hat{\boldsymbol{h}}_{N_t}^T]^T$ denotes the quantized channels of the strong users, where S_{su} denotes the selection set of the strong users. The BS performs ZFBF based on $\hat{\boldsymbol{H}}(S_{su})$ and designs the unnormalized precoding matrix, $\boldsymbol{W}_0$, as

$$\begin{aligned} \boldsymbol{W}_0(S_{su}) &= \hat{\boldsymbol{H}}(S_{su})^* \left(\hat{\boldsymbol{H}}(S_{su})\hat{\boldsymbol{H}}(S_{su})^*\right)^{-1} \\ &= [\boldsymbol{w}_{0_1}, \ldots, \boldsymbol{w}_{0_{N_t}}], \end{aligned} \tag{7}$$

Each kth column of $\boldsymbol{W}_0(S_{su})$ is then normalized as $\boldsymbol{w}_k = \boldsymbol{w}_{0_k} / \|\boldsymbol{w}_{0_k}\|$ to obtain the normalized precoding matrix

$$\boldsymbol{W}(S_{su}) = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{N_t}], \tag{8}$$

Since the precoding matrix $\boldsymbol{W}(S_{su})$ is designed based on the quantized CDIs of the strong users, $\hat{\boldsymbol{H}}(S_{su})$ and not the actual channels $\boldsymbol{H}(S_{su})$, the multi-user interference cannot be completely eliminated, and this leads to performance degradation compared to the case of full CSI.

Denote that under partial CSI with ZFBF, the effective channel gain of the strong user $k1$ can be computed by solving the inverse of the $(k1, k1)$ element of $\left(\hat{\boldsymbol{H}}(S_{su})\hat{\boldsymbol{H}}(S_{su})^*\right)^{-1}$ matrix as follows:

$$\lambda_{k1} = \frac{1}{\|\boldsymbol{w}_{0_{k1}}\|^2} = \frac{1}{\left[\left(\hat{\boldsymbol{H}}(S_{su})\hat{\boldsymbol{H}}(S_{su})^*\right)^{-1}\right]_{k1,k1}} \tag{9}$$

III. SIGNAL MODEL FOR NOMA-ZFBF USERS

The received superimposed signal in downlink NOMA-ZFBF contains two types of interferences: the first is the inter-cluster interference (ICI) which comes from users of other clusters and can be removed by ZF equalizer at the receiver side. The second is the inter-user interference (IUI) which caused by the users within the same cluster. To extract the strong user symbol from the superimposed signal, the strong user subtracts the IUI comes from weak user's signal using SIC, then decodes its own message. The weak user can directly detect its symbol without performing SIC since the BS allocates more power to the weak user to increase its signal to interference plus noise ratio (SINR) for a better detection. In the following subsections, the signal model, data rate and other detection details are presented for strong and weak users.

A. SIGNAL MODEL AND DATA RATE: STRONG USER

By substituting (1) in (2), we can rewrite the received signal of the strong user before performing SIC and ZFBF as:

$$\begin{aligned} y_{k1} = &\underbrace{\sqrt{\alpha_k P}\boldsymbol{h}_{k1}\boldsymbol{w}_k s_{k1}}_{\text{Desired signal}} + \underbrace{\sqrt{(1-\alpha_k)P}\boldsymbol{h}_{k1}\boldsymbol{w}_k s_{k2}}_{\text{Inter-user interference}} \\ &+ \underbrace{\sqrt{P}\sum_{j=1, j\neq k}^{N_t} \boldsymbol{h}_{k1}\boldsymbol{w}_j x_j}_{\text{Inter-cluster interference}} + \underbrace{n_{k1}}_{\text{Noise}}. \end{aligned} \tag{10}$$

After performing SIC and ZF equalization, IUI can be completely removed, while the ICI cannot be totally eliminated due to channel quantization error. Therefore, the received signal is reduced to:

$$y_{k1} = \sqrt{\alpha_k P}\boldsymbol{h}_{k1}\boldsymbol{w}_k s_{k1} + \sqrt{P}\sum_{j=1, j\neq k}^{N_t} \boldsymbol{h}_{k1}\boldsymbol{w}_j x_j + n_{k1}. \tag{11}$$

And accordingly, the SINR of the strong user is given by:

$$\begin{aligned} SINR_{k1} &= \frac{\alpha_k P |\boldsymbol{h}_{k1}\boldsymbol{w}_k|^2}{P\sum_{j=1, j\neq k}^{N_t} |\boldsymbol{h}_{k1}\boldsymbol{w}_j|^2 + \sigma_n^2} \\ &= \frac{\alpha_k \rho |\boldsymbol{h}_{k1}\boldsymbol{w}_k|^2}{\rho\sum_{j=1, j\neq k}^{N_t} |\boldsymbol{h}_{k1}\boldsymbol{w}_j|^2 + 1} \end{aligned} \tag{12}$$

where $\rho = P/\sigma_n^2$ is the transmit SNR of the kth cluster. If the channel quality indicator (CQI) of the strong user is defined as

$$\gamma_{k1} = \frac{\rho |\boldsymbol{h}_{k1}\boldsymbol{w}_k|^2}{\rho\sum_{j=1, j\neq k}^{N_t} |\boldsymbol{h}_{k1}\boldsymbol{w}_j|^2 + 1} \tag{13}$$

The achievable sum capacity of the strong user can be expressed in terms of CQI as:

$$R_{k1} = \log\left(1 + \alpha_k \gamma_{k1}\right). \tag{14}$$

B. SIGNAL MODEL AND DATA RATE: WEAK USER

Since the weak user does not perform the SIC process, the IUI is not eliminated from the received signal. Also, the ICI cannot be completely removed because the beams are designed based on strong user channels. As a result, the received signal of the weak user has both IUI and ICI and can be described as

$$y_{k2} = \sqrt{(1-\alpha_k)P}\boldsymbol{h}_{k2}\boldsymbol{w}_k s_{k2} + \sqrt{\alpha_k P}\boldsymbol{h}_{k2}\boldsymbol{w}_k s_{k1} + \sqrt{P}\sum_{j=1, j\neq k}^{N_t} \boldsymbol{h}_{k2}\boldsymbol{w}_j x_j + n_{k2}. \tag{15}$$

Accordingly, the SINR for the weak user contains both IUI and ICI terms and is given by:

$$SINR_{k2} = \frac{(1-\alpha_k)\, P\, |\boldsymbol{h}_{k2}\boldsymbol{w}_k|^2}{\alpha_k P\, |\boldsymbol{h}_{k2}\boldsymbol{w}_k|^2 + P\sum_{j=1, j\neq k}^{N_t} \left|\boldsymbol{h}_{k2}\boldsymbol{w}_j\right|^2 + \sigma_n^2} = \frac{(1-\alpha_k)\, \rho\, |\boldsymbol{h}_{k2}\boldsymbol{w}_k|^2}{\alpha_k \rho\, |\boldsymbol{h}_{k2}\boldsymbol{w}_k|^2 + \rho\sum_{j=1, j\neq k}^{N_t} \left|\boldsymbol{h}_{k2}\boldsymbol{w}_j\right|^2 + 1} \tag{16}$$

If the is the CQI of the weak user is defined as

$$\gamma_{k2} = \frac{\rho\, |\boldsymbol{h}_{k2}\boldsymbol{w}_k|^2}{\rho\sum_{j=1, j\neq k}^{N_t} \left|\boldsymbol{h}_{k2}\boldsymbol{w}_j\right|^2 + 1} \tag{17}$$

Then, the data rate of the weak user R_{k2} in the kth cluster can be expressed as:

$$R_{k2} = \log\left(1 + \frac{(1-\alpha_k)\,\gamma_{k2}}{\alpha_k \gamma_{k2} + 1}\right). \tag{18}$$

IV. CQI FEEDBACK MODEL

In partial CSI feedback, the users only provide their quantized CDI and CQI values, which are used by the BS to schedule users and to determine the ZFBF. To derive the expected CQI, we start with the strong users and then generalize it for all users. Let $\tilde{\boldsymbol{h}}_{k1} = \boldsymbol{h}_{k1}/\|\boldsymbol{h}_{k1}\|$ denotes the direction of the strong user channel. Therefore, we can rewrite (13) in terms of $\tilde{\boldsymbol{h}}_{k1}$ as

$$\gamma_{k1} = \frac{\rho\, \|\boldsymbol{h}_{k1}\|^2\, |\tilde{\boldsymbol{h}}_{k1}\boldsymbol{w}_k|^2}{\rho\, \|\boldsymbol{h}_{k1}\|^2 \sum_{j=1, j\neq k}^{N_t} |\tilde{\boldsymbol{h}}_{k1}\boldsymbol{w}_j|^2 + 1} \tag{19}$$

According to RVQ, the true CDI $\tilde{\boldsymbol{h}}_{k1}$ can be decomposed as [18], [20]

$$\tilde{\boldsymbol{h}}_{k1} = \cos\theta_{k1}\hat{\boldsymbol{h}}_{k1} + \sin\theta_{k1}\tilde{\boldsymbol{e}}_{k1}, \tag{20}$$

where $\theta_{k1} \in [0, \pi/2]$ is the angle the angle between the real channel $\boldsymbol{h}_{k1}$ and its quantization $\hat{\boldsymbol{h}}_{k1}$, i.e., $\cos\theta_{k1} = \left|\tilde{\boldsymbol{h}}_{k1}\hat{\boldsymbol{h}}_{k1}^*\right|$, and $\tilde{\boldsymbol{e}}_{k1} = \boldsymbol{e}_{k1}/\|\boldsymbol{e}_{k1}\|$ is an error vector due to channel quantization.

Substituting (20) in the denominator of (19), and applying (6), i.e., $\hat{\boldsymbol{h}}_{k1}\boldsymbol{w}_j = 0$ for $k \neq j$. The inner product of $\hat{\boldsymbol{h}}_{k1}$ and $\boldsymbol{w}_j$ is then given by

$$|\tilde{\boldsymbol{h}}_{k1}\boldsymbol{w}_j|^2 = \cos^2\theta_{k1}|\hat{\boldsymbol{h}}_{k1}\boldsymbol{w}_j|^2 + \sin^2\theta_{k1}|\tilde{\boldsymbol{e}}_{k1}\boldsymbol{w}_j|^2 = \sin^2\theta_{k1}|\tilde{\boldsymbol{e}}_{k1}\boldsymbol{w}_j|^2 \tag{21}$$

Since both $\tilde{\boldsymbol{e}}_{k1}$ and $\boldsymbol{w}_j$ are i.i.d. isotropic vectors on the $(N_t - 1)$ dimensional nullspace of $\hat{\boldsymbol{h}}_{k1}$, the quantity $|\tilde{\boldsymbol{e}}_{k1}\boldsymbol{w}_j|^2$ is a Beta $(1, N_t - 2)$ random variable (i.e., $|\tilde{\boldsymbol{e}}_{k1}\boldsymbol{w}_j|^2 = \beta(1, N_t - 2)$) and independent on θ_{k1}. Therefore, the expectation of $|\tilde{\boldsymbol{h}}_{k1}\boldsymbol{w}_k|^2$ can be evaluated as

$$\mathbb{E}\left[|\tilde{\boldsymbol{h}}_{k1}\boldsymbol{w}_k|^2\right] \leq \sin^2\theta_{k1}\,\mathbb{E}\left[\beta(1, N_t - 2)\right] \tag{22}$$

where the inequality follows Jensen's inequality. Since a beta random variable limitation is between [0,1], we clearly have $|\tilde{\boldsymbol{h}}_{k1}\boldsymbol{w}_k|^2 = \sin^2\theta_{k1}$. This also means that the interference from any user is no larger than the quantization error and no beta random variable is needed [21].

Therefore, the quantity $|\tilde{\boldsymbol{h}}_{k1}\boldsymbol{w}_k|^2$ in the numerator of (19) can be calculated as

$$|\tilde{\boldsymbol{h}}_{k1}\boldsymbol{w}_k|^2 = \cos^2\theta_{k1}|\hat{\boldsymbol{h}}_{k1}\boldsymbol{w}_k|^2 + \sin^2\theta_{k1}|\tilde{\boldsymbol{e}}_{k1}\boldsymbol{w}_k|^2 = \cos^2\theta_{k1} \tag{23}$$

Since the beamforming direction $\boldsymbol{w}_k$ is designed to be closely aligned with $\hat{\boldsymbol{h}}_{k1}$ for $k = j$ and orthogonal to $\tilde{\boldsymbol{e}}_{k1}$ for $k \neq j$, we have, $\hat{\boldsymbol{h}}_{k1}\boldsymbol{w}_k \approx 1$ and $\tilde{\boldsymbol{e}}_{k1}\boldsymbol{w}_k \approx 0$.

Then, the expected SINR for the kth strong user becomes

$$\mathbb{E}\left[\gamma_{k1}\right] = \mathbb{E}\left[\frac{\rho\, \|\boldsymbol{h}_{k1}\|^2\, |\tilde{\boldsymbol{h}}_{k1}\boldsymbol{w}_k|^2}{\rho\, \|\boldsymbol{h}_{k1}\|^2 \sum_{j=1, j\neq k}^{N_t} |\tilde{\boldsymbol{h}}_{k1}\boldsymbol{w}_j|^2 + 1}\right] = \frac{\rho\, \|\boldsymbol{h}_{k1}\|^2 \cos^2\theta_{k1}}{1 + \rho\, \|\boldsymbol{h}_{k1}\|^2 \sin^2\theta_{k1}} = \hat{\gamma}_{k1} \tag{24}$$

In general, for any user k, the feedback CQI is given by:

$$\hat{\gamma}_k = \frac{\rho\, \|\boldsymbol{h}_k\|^2 \cos^2\theta_k}{1 + \rho\, \|\boldsymbol{h}_k\|^2 \sin^2\theta_k} = \rho\|\hat{\boldsymbol{h}}_k\|^2 \tag{25}$$

which reflects the channel norm and the quantization error, that the transmitter must have in the quantized CDI feedback scheme, rather than that applied for full CSI schemes, i.e., $\rho\|\boldsymbol{h}_k\|^2$.

A. GENERATION OF NUMERICAL RESULTS

As the number of feedback bits B became larger than 15 or 20 bits, the numerical calculations of $\hat{\boldsymbol{h}}_k$ based on (3) become computationally complex. Therefore, we present in Algorithm 1 an RVQ codebook generator, which can be exploited to model an efficient and precise quantized channel. The first step in the for loop generates an error for each user k based on RVQ. The second step, calculates the normalized feedback channel with the generated error form step 1. The last step, is to find the quantized channel by interchanging $\tilde{\boldsymbol{h}}_k$ and $\hat{\boldsymbol{h}}_k$, in (20), which yields an equivalent decomposition. The resulted channel is then multiplied by $\sqrt{\hat{\gamma}_k/\rho}$ to remove the effect of normalization and ρ and to give the final quantized channel vector $\hat{\boldsymbol{h}}_k$.

Algorithm 1 Quantized Channel Model With RVQ Codebook Generator

for $\forall k \in \{1, \ldots, K\}$ **do**

1) Find the minimum quantization error angle θ_k for user k such that

$$\mathbb{E}\left[\sin^2 \theta_k\right] \leq 2^{-\frac{B}{N_t-1}} \tag{26}$$

2) Calculate the CQI $\hat{\gamma}_k$ for user k according to (25).
3) Compute $\hat{\boldsymbol{h}}_k$, using the following decomposition

$$\hat{\boldsymbol{h}}_k = \sqrt{\frac{\hat{\gamma}_k}{\rho}} \left(\cos\theta_k \tilde{\boldsymbol{h}}_k + \sin\theta_k \tilde{\boldsymbol{e}}_k\right), \tag{27}$$

end for

V. PROPORTIONAL FAIR SCHEDULING

Proportional Fair (PF) is a compromise-based scheduling scheme that can be used to maximize the total network throughput while at the same time, guarantees user fairness, by considering the users' instantaneous channel qualities and their average throughputs. Therefore, it has been widely adopted in different wireless communication systems to schedule downlink data flow among different users. In multiuser transmission scenario, the BS schedules a set of users according to PF selection metric as [14]

$$S = \arg\max_{S \in U} \sum_{i \in S} \frac{R_i(t)}{\mu_i(t)}, \tag{28}$$

where $R_i(t)$ and $\mu_i(t)$ are the current data rate and the average data rate, respectively, received by the user i, at time slot t and S is a set of scheduled users. The average data rate $\mu_i(t)$ is updated at every time slot, according to [14]:

$$\mu_i(t+1) = \begin{cases} \left(1 - \frac{1}{t_c}\right)\mu_i(t) + \frac{1}{t_c} R_i(S,t), & i \in S, \\ \left(1 - \frac{1}{t_c}\right)\mu_i(t), & i \notin S, \end{cases} \tag{29}$$

where t_c is the time window size which is settled to maintain fairness over a pre-determined time horizon. The value of t_c can highly determine the trade-off between the throughput and fairness, i.e. with small t_c values, a high level of fairness can be satisfied at the cost of decreasing the throughput level and vice versa. Therefore, the t_c value must be appropriately chosen in order to improve this trade-off.

A. FAIRNESS: CONCEPTS AND PERFORMANCE METRIC

In wireless networks, the user fairness indicates how equally radio resources are distributed between mobile users. The result of an unfair share of system resources may lead to decrease the required QoS among some users, or redundant allocation for others.

To measure fairness, Jain's fairness index (JFI) (or simply fairness index) [22] is used to evaluate the level of fairness achieved over a finite horizon. The index has a range of [0,1] to allow for comparison. If all the users get the same amount of resources (i.e., all are equal), then JFI is 1, and the system is 100% fair. As the inequlity of resurces increases fairness is decreases and JFI will be near 0. Jain's index is defined in [22], [23] as

$$J = \frac{\left(\sum_{i=1}^{K} T_i\right)^2}{K \sum_{i=1}^{K} T_i^2}, \tag{30}$$

where K denoting the number of considered users, T_i is the average throughput allocated to user i and calculated over a finite time horizon of length W_t.

VI. TRANSMIT POWER ALLOCATION

Power allocation (PA) is a vital component in wireless communication which is used to control the interference, that's directly effects on the throughput. In NOMA system, the users are multiplexed in the power domain. Thus, the power allocated to one user affects the achievable throughput of not only that user but also that of other users. Most of the existing power allocation algorithms focused on enhancing the total throughput and neglecting weak user throughput. Therefore, in this work, the power allocation optimization problems is formulated to maximize the total throughput without compromising the weak users' sum capacity under both full and partial CSI. By allocating the required power amount that ensures achieving the same or even higher throughput of the weak if its works in time division multiple access (TDMA) system. Thus, the optimization formula is defined as follows:

$$\max_{\alpha_k} \; (R_{k1} + R_{k2}) \tag{31a}$$

$$s.t.\; R_{k2} = \frac{1}{2} R_{k2-TDMA} \tag{31b}$$

$$R_{k1} = \frac{1}{2} R_{k1-TDMA} \tag{31c}$$

$$0 \leq \alpha_k \leq 1 \tag{31d}$$

$$(1 - \alpha_k) \geq \alpha_k \tag{31e}$$

In this optimization problem, the $1/2$ value in constraint (31b) and (31c) is used to make a balance between TDMA and NOMA systems, because the TDMA system needs two-time slots to serve $2N_t$ users that can be served in a single time slot t with NOMA system. Constraint (31e) ensures allocation more power to the weak user than that of the strong user.

The problem defined in (31) is convex with respect to α_k and its Karush-Kuhn-Tucker (KKT) conditions are given as follows:

$$(i)\;\; \frac{\partial (R_{k1} + R_{k2})}{\partial \alpha_k} = \Lambda \frac{\partial \left(R_{k2} - \frac{1}{2} R_{k2-TDMA}\right)}{\partial \alpha_k},$$

$$(ii)\;\; \Lambda \geq 0,$$

$$(iii)\;\; R_{k2} - \frac{1}{2} R_{k2-TDMA} \leq 0,$$

$$(iv)\;\; \Lambda \left(R_{k2} - \frac{1}{2} R_{k2-TDMA}\right) = 0, \tag{32}$$

where Λ is the Lagrange multipliers for the constraints. Clearly, $\Lambda \neq 0$. Otherwise, $\alpha_k < 0$ cannot satisfy (31d). Therefore, we can solve conditions (*iv*) in (32) for the optimal solution as:

$$R_{k2} = \frac{1}{2} R_{k2-TDMA}, \tag{33}$$

$$\log\left(1 + \frac{(1-\alpha_k)\,\gamma_{k2}}{\alpha_k \gamma_{k2} + 1}\right) = \frac{1}{2}\log\left(1 + SINR_{k2-TDMA}\right) \tag{34}$$

where the corresponding sum-rates for the selected weak users in the TDMA system is given by

$$R_{k2-TDMA} = \log\left(1 + SINR_{k2-TDMA}\right), \tag{35}$$

and

$$SINR_{k2-TDMA} = \frac{\rho\,|\boldsymbol{h}_{k2}\boldsymbol{w}_k|^2}{\rho\sum_{j=1,j\neq k}^{N_t}|\boldsymbol{h}_{k2}\boldsymbol{w}_j|^2 + 1} \tag{36}$$

Hence, the optimal power fraction of the strong user α_k^* is obtained from (34), and its given by:

$$\alpha_k^* = \frac{1 + \frac{1}{\gamma_{k2}}}{\sqrt{1 + SINR_{k2-TDMA}}} - \frac{1}{\gamma_{k2}} \tag{37}$$

And accordingly, $(1 - \alpha_k^*)$ power fraction is allotted to the weak user. Note that if α_k^* is outside the two feasible regions i.e., $\alpha_k^* \notin [0, 1]$, the NOMA-BF system fails to achieve a better data rate performance than that of the TDMA-BF system. Therefore, the kth cluster is switched to the conventional TDMA-BF transmission instead of NOMA-BF and only the strong user will be supported by the BS. It's worth to mention that the proposed power allocation α_k^* is different from those proposed in [11], [12], which are suitable for full CSI case only, whereas our proposed α_k^* is extended for both full and partial CSI cases.

VII. PROPOSED ALGORITHMS

User scheduling is one of the key techniques used to increase the spectral efficiency in wireless systems by exploiting the multiuser diversity gain. Most of the existing multiuser NOMA scheduling schemes only focused on enhancing the total system capacity, and therefore, the QoS is dropped for some users especially the weak users, which may not be selected for a long time, due to their low channel quality. Therefore, we propose PF-MPECG-SIR and PF-MPECG-CORR algorithms based on PF, to make the user selection depends on the average throughput in addition to the channel quality to achieve a better QoS. Each of the proposed algorithms has two stages, as described in Algorithm 2 and Algorithm 3. In the first stage, the BS selects the strong users, then, searching for the weak users in the second stage to form a "cluster", in which NOMA principle is applied. The proposed algorithms are described in detail in the following subsections.

Algorithm 2 Proposed PF-MPECG-SIR Algorithm

Stage 1: Strong Users Selection: PF-MPECG

1: At time t, initialize: $U = \{1, \ldots, K\}$, $S_{su} = \varnothing$, $\mu(1) = 1$ for all users, and set $k = 1$ to select the first strong user according to

$$S_{su}(1) = \arg\max_{i\in U}\left(\frac{\log(1 + P\|\hat{\boldsymbol{h}}_i\|^2)}{\mu_i(t)}\right), \tag{38}$$

2: $S_{su} \leftarrow S_{su} \cup \{S_{su}(k)\}$ and $k \leftarrow k + 1$

3: Find the kth strong user, from the remaining user set $C = U - S_{su}$, after setting the CDI matrix $\hat{\boldsymbol{H}}(S_{su})$ as

$$\hat{\boldsymbol{H}}(S_{su}) = \left[\hat{\boldsymbol{H}}(S_{su}(k-1))^T, \hat{\boldsymbol{h}}_i^T\right]^T, i \in C \tag{39}$$

$$S_{su}(k) = \arg\max_{i\in C}\left(\frac{\log\left(1 + \prod_{j=1}^{k}\lambda_j\right)}{\mu_i(t)}\right), \tag{40}$$

where

$$\lambda_j = \frac{1}{\left[\hat{\boldsymbol{H}}(S_{su})\,\hat{\boldsymbol{H}}(S_{su})^*\right]_{j,j}^{-1}}, \tag{41}$$

4: After selecting the kth strong user, update $\hat{\boldsymbol{H}}(S_{su})$ as

$$\hat{\boldsymbol{H}}(S_{su}) = \left[\hat{\boldsymbol{H}}(S_{su}(k-1))^T, \hat{\boldsymbol{h}}_k^T\right]^T, \tag{42}$$

5: $S_{su} \leftarrow S_{su} \cup \{S_{su}(k)\}$ and $k \leftarrow k + 1$

6: If $k \leq N_t$, go to step 3. Else, strong users selection is completed.

7: Generate ZFBF matrix $\boldsymbol{W}(S_{su})$ using (8) to be used in the selection of the weak users and sum-rate calculations.

Stage 2: Weak Users Selection: PF-SIR

8: Initialize: $S_{wu} = \varnothing$ and set $k = 1$.

9: Search within $C = U - S_{su} - S_{wu}$ to select the weak user for cluster k as

$$S_{wu}(k) = \arg\max_{i\in C}\left(\frac{\log\left(1 + \frac{|\hat{\boldsymbol{h}}_i\boldsymbol{w}_k|^2}{\sum_{j=1,j\neq k}^{N_t}|\hat{\boldsymbol{h}}_i\boldsymbol{w}_j|^2}\right)}{\mu_i(t)}\right), \tag{43}$$

10: $S_{wu} \leftarrow S_{wu} \cup \{S_{wu}(k)\}$ and $k \leftarrow k + 1$

11: If $k \leq N_t$, go back to 9. Else, weak users selection is completed.

12: Find the optimal power division between the selected strong and weak users in each cluster using (37).

13: Calculate the data rates for the strong and weak users using (14) and (18), respectively.

14: Update $\mu_i(t + 1)$ for all users using (29).

15: Go to the strong user selection stage to make new clusters for the next time slot $(t + 1)$.

A. STRONG USERS SELECTION BY PF-MPECG

To achieve high data rates for the strong users and guarantee fairness among users, the PF scheduling algorithm is

Algorithm 3 Proposed PF-CORR to Select the Weak Users

1: Initialize: $S_{wu} = \varnothing$ and set $k = 1$.
2: Search within $C = U - S_{su} - S_{wu}$ to find the kth weak user which can maximize the following selection criterion

$$S_{wu}(k) = \arg\max_{i \in C} \left(\frac{Corr\left(\hat{\boldsymbol{h}}_k, \hat{\boldsymbol{h}}_i\right)}{\mu_i(t)} \right), \tag{44}$$

where

$$Corr\left(\hat{\boldsymbol{h}}_k, \hat{\boldsymbol{h}}_i\right) = \frac{|\hat{\boldsymbol{h}}_k \hat{\boldsymbol{h}}_i^*|}{\|\hat{\boldsymbol{h}}_k\| \|\hat{\boldsymbol{h}}_i\|}, \tag{45}$$

3: $S_{wu} \leftarrow S_{wu} \cup \{S_{wu}(k)\}$ and $k \leftarrow k + 1$
4: If $k \leq N_t$, go to step 2. Otherwise, weak users scheduling is completed.
5: Follow steps 12 to 15 in Algorithm 2.

modified and extended to include the maximum product of effective channel gain (PF-MPECG). As shown in Algorithm 2, the first strong user is chosen from the initial user set U based on the largest quantized channel gain to the average throughput $\mu_i(t)$ ratio, given in (38). While the other joined strong users' should be selected according to (40) which selects the kth user that can maximize the product of effective quantized channel gains to the average throughput $\mu_i(t)$ ratio. After selecting N_t strong users, the BS designs the ZF beams that used to choose the weak users in the second stage.

B. WEAK USERS SELECTION BY PF-SIR

In this stage, weak users are selected from the remaining $C = U - S_{su}$ users, using the designed beams from the strong users' CDIs. With the PF-SIR scheme, which summarized in the second stage of Algorithm 2, the kth weak user is selected by maximizing the signal to interference to average user throughput $\mu_i(t)$ ratio, which is given in (43). After selection N_t weak users, the BS allocates the optimal power portion to the strong and weak users, according to (37). Finally, all users average throughputs $\mu_i(t+1)$ are updated using (29) to be used in the next time slot t.

C. WEAK USERS SELECTION BY PF-CORR

If the correlation is integrated into the PF selection principle, the fairness can be highly enhanced with a slight decrease in the total system sum-rate. The PF-CORR is summarized in Algorithm 3. According to this algorithm, the BS selects the kth weak user with the maximum correlation to average user throughput $\mu_i(t)$ ratio, which is given in (44). The selection process is repeated until N_t weak users are selected. The remaining steps follow the steps 12 to 15 described in Algorithm 2.

VIII. COMPUTATIONAL COMPLEXITY ANALYSIS

In this section, we quantify the complexity of the proposed PF-MPECG-SIR, and the PF-MPECG-CORR algorithms. The complexity can be counted as the number of flops, where a flop is defined as a real floating point operation and denoted as ψ. In the complexity calculations, we adopt the following counting rules:

- A real addition, a multiplication, and a division are counted as: 1 flop.
- A complex addition and multiplication are considered as two flops and six flops, respectively.
- Multiplication of two $N_t \times N_t$ complex matrices, has: $8\,N_t^3$ flops.
- An inversion of $N_t \times N_t$ complex matrix needs: $\frac{8}{3}N_t^3$ flops [24].

A. COMPLEXITY OF PF-MPECG-SIR ALGORITHM

The total complexity of the PF-MPECG-SIR algorithm is given by:

1) The squared Hilbert-Schmidt norm for each user CDI $\|\hat{\boldsymbol{h}}_i\|^2$, requires $(1 \times N_t) \times (N_t \times 1)$ vector-matrix multiplications, which results in $2N_t$ real multiplications $+ 2N_t - 1$ real additions. For K users, $2KN_t$ real multiplications and $K(2N_t - 1)$ real summations are required. The division by $\mu_i(t)$, required a complexity of $\mathcal{O}(K)$ for K users, Therefore, the flop count in step 1, is $4KN_t - K + K = 4KN_t$.
2) To count $\left[\hat{\boldsymbol{H}}(S_{su})\,\hat{\boldsymbol{H}}(S_{su})^*\right]^{-1}_{j,j}$ complexity in step 3. Each user needs: A complexity order of $\mathcal{O}(8N_t^3)$ for one complex matrices multiplication and $\mathcal{O}(\frac{8}{3}N_t^3)$ for $(N_t \times N_t)$ complex matrix inversion. In addition to $\mathcal{O}(1)$, for the division by $\mu_i(t)$. Therefore, for $K - 1$ users, the total complexity of step 3, is of $\mathcal{O}\left(\frac{32}{3}(K-1)N_t^3 + K - 1\right)$.
3) In step 7, the calculation of ZFBF, takes two complex matrices multiplications, one complex matrix inversion and N_t column normalization for the resultant matrix. Two matrix multiplications with $(N_t \times N_t)$ dimension needs a complexity of $\mathcal{O}(16N_t^3)$, matrix inversion needs $\mathcal{O}(\frac{8}{3}N_t^3)$ and to normalize N_t columns, it takes, $4N_t - 1$ flops to normalize each column. In total, step 7, needs a complexity of $\mathcal{O}(\frac{56}{3}N_t^3 + 4N_t^2 - N_t)$.
4) In step 9, to calculate the PF-SIR for the remaining $K - N_t$ users, we need for each user: $(1 \times N_t) \times (N_t \times 1)$ multiplication for $|\hat{\boldsymbol{h}}_i \boldsymbol{w}_k|^2$ term, which requires $4N_t - 1$ flops. The $N_t - 1$ interference terms, also requires $4N_t - 1$ flops for each term, in addition to $N_t - 1$ real additions to sum these interferences, and two real divisions are needed, the first division is to divide the signal power to interference and the second one is to divide by $\mu_i(t)$. Thus, for $K - N_t$ users, the flop count in step 9 is: $[4N_t - 1 + (N_t - 1)(4N_t - 1) + N_t - 1 + 2](K - N_t) = 4N_t^2K + K - 4N_t^3 - N_t$.

The complexity of the remaining steps, i.e., power allocation, computing the users data rates and updating $\mu_i(t+1)$ is not included in the complexity analysis, as they are not related to the user selection procedures. Hence, the total flop count of

the PF-MPECG-SIR algorithm becomes:

$$\begin{aligned}\psi_{PF-MPECG-SIR} &= 4KN_t + \frac{32}{3}(K-1)N_t^3 + K - 1 \\ &\quad + \frac{56}{3}N_t^3 + 4N_t^2 - N_t + 4N_t^2K + K \\ &\quad - 4N_t^3 - N_t \\ &\approx \frac{32}{3}(K-1)N_t^3 + 4N_t^2K \approx \mathcal{O}(KN_t^3)\end{aligned} \tag{46}$$

B. COMPLEXITY OF PF-MPECG-CORR ALGORITHM

When weak users are selected based on the PF-CORR, the complexity of the second stage will be:

- For $|\hat{\boldsymbol{h}}_k\hat{\boldsymbol{h}}_i^*|$ term, we need $(1 \times N_t) \times (N_t \times 1)$ vector-matrix multiplication, which takes $4N_t - 1$ flops.
- Normalizing of $\|\hat{\boldsymbol{h}}_k\|\|\hat{\boldsymbol{h}}_i\|$ terms, require $2 \times (4N_t - 1) = 8N_t - 2$ flops, in addition to one flop to multiply both of them. thus, this step we need, $8N_t - 1$ flops.
- Three flops are required, two real divisions and one real multiplication.

Therefore, the complexity of PF-CORR for $K - N_t$ users will be:

$$\begin{aligned}\psi_{PF-CORR} &= (4N_t - 1 + 8N_t - 1 + 3)(K - N_t) \\ &= (12N_t + 1)(K - N_t) \\ &= 12KN_t - 12N_t^2 + K - N_t\end{aligned} \tag{47}$$

Thus, the total flop count of the PF-MPECG-CORR algorithm becomes:

$$\begin{aligned}\psi_{PF-MPECG-CORR} &= 4KN_t + \frac{32}{3}(K-1)N_t^3 + K - 1 \\ &\quad + \frac{56}{3}N_t^3 + 4N_t^2 - N_t + 12KN_t \\ &\quad - 12N_t^2 + K - N_t \\ &\approx \frac{32}{3}(K-1)N_t^3 + 12KN_t \approx \mathcal{O}(KN_t^3)\end{aligned} \tag{48}$$

Therefore, the computational complexity for the PF-MPECG-SIR and PF-MPECG-CORR algorithms is proportional to $\mathcal{O}(KN_t^3)$, which is the same as the MPECG [19] and SUS [13] user selection algorithms.

IX. SIMULATION RESULTS

In this section, the performance of the proposed PF-MPECG-SIR and PF-MPECG-CORR algorithms under full and partial CSI is presented. The number of the BS antennas is $N_t = 2$, the total power for each cluster $P = 15$ dB, the channel and noise parameters are described in the system model. Different numbers of candidate users' K are used in the simulation. However, the simulation tests focused on $K = 150$, to measure the achieved throughput-fairness trade-off of our proposed system over the conventional SUS-SIR [12].

In Fig. 2, the sum-rate performance of MPECG-SIR, PF-MPECG-SIR and PF-MPECG-CORR are compared with

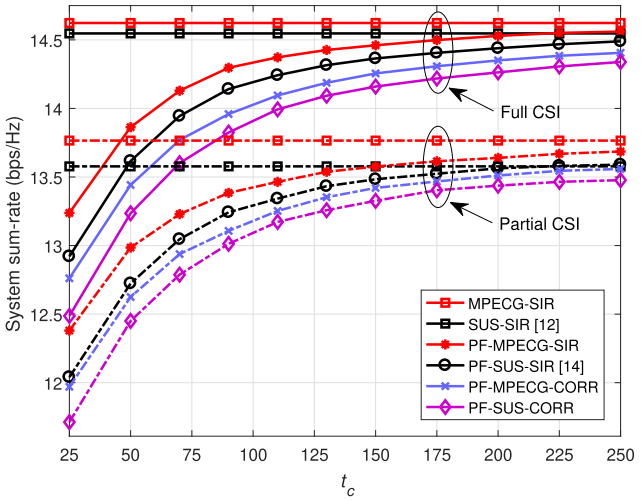

FIGURE 2. Total system sum-rate performance for different user clustering algorithms versus t_c. The number of users $K = 150$ users. For partial CSI, the number of feedback bits $B = 8$ bits/user.

the SUS-SIR [12], PF-SUS-SIR [14] and the PF-SUS-CORR with respect to different t_c values in case of full and partial CSI. It can be seen that the SUS-SIR and the MPECG-SIR throughputs are not changed for all t_c values since they are not PF based schemes. However, we can clearly notice that the MPECG-SIR can achieve a better throughput compared to the SUS-SIR, especially, in partial CSI. We notice the same behavior for these schemes when modified to be PF based, i.e., the PF-MPECG-SIR and PF-MPECG-CORR schemes outperform the PF-SUS-SIR and the PF-SUS-CORR schemes for both full and partial CSI. These comparisons prove the superiority of our proposed scheme over the traditional SUS-SIR scheme, especially in the partial feedback scenario.

Moreover, we observe at $t_c = 225$, the PF-MPECG-SIR can achieve the same throughput as the SUS-SIR, in case of full CSI scenario, and surpass it, in case of partial CSI. Therefore, in this work, all the rest of the simulation results are based on $t_c = 225$ in which the SUS-SIR throughput can be achieved, and user fairness can be improved.

Fig. 3-Fig. 4, show the average throughput performance of the total system and weak user versus a different number of users K for PF-MPECG-SIR, PF-MPECG-CORR algorithms and compare them with different SUS-SIR algorithms [12], [18] for both full and partial CSI. We can see in Fig. 3 that the performance of all techniques increases as the number of candidate users K increases due to multiuser diversity gain. In addition, we observe that all NOMA-ZFBF schemes give higher sum-rates than that of TDMA-ZFBF schemes. Moreover, it can be seen that the PF-MPECG-SIR gives the same performance as SUS-SIR [12], while PF-MPECG-CORR has a slightly lower throughput performance since it is basically designed to maximize user fairness. However, this difference is diminished in partial CSI condition, as shown in Fig. 3. In addition, we noticed that the total system performance of SUS-SIR [18] can give slightly

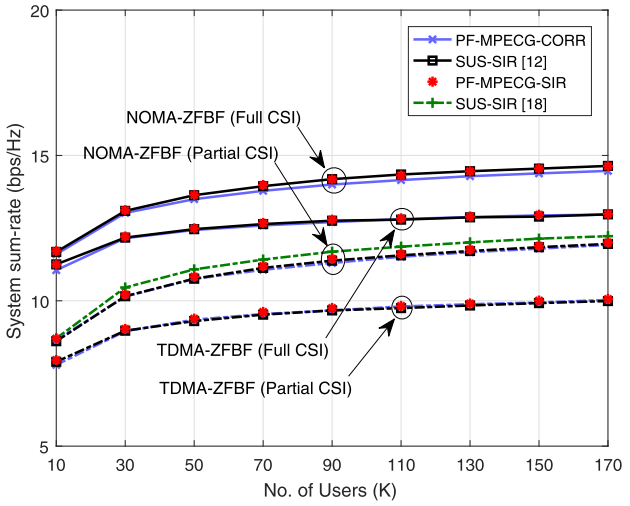

FIGURE 3. Total system sum-rate performance versus the number of users. Solid curves refer to the full CSI and the dashed curves refer to the partial CSI with $B = 4$ bits/user.

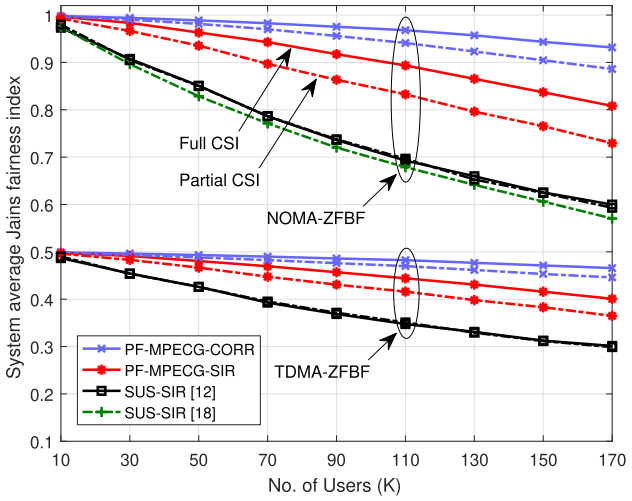

FIGURE 5. Jain's fairness index of the overall system versus the number of users. Solid curves refer to the full CSI. dashed curves refer to the partial CSI with $B = 4$ bits/user.

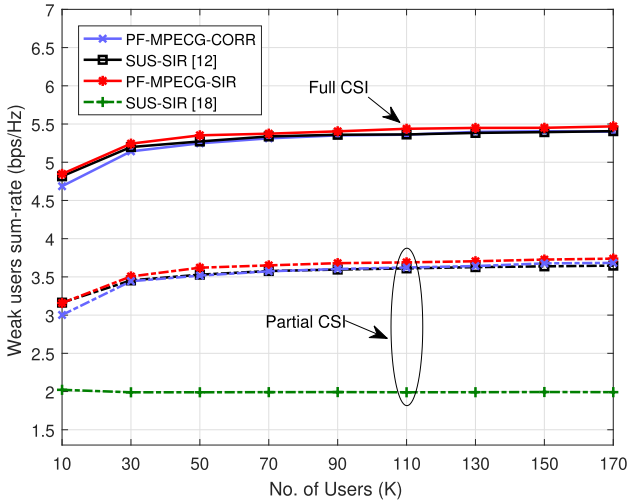

FIGURE 4. Weak users' sum-rate versus the number of users for NOMA-ZFBF and TDMA-ZFBF systems. Both systems have identical sum-rates for all the schemes except the SUS-SIR [18], which represents NOMA-ZFBF sum-rate only. Solid curves refer to the full CSI and the dashed curves refer to the partial CSI with $B = 4$ bits/user.

higher throughput than our proposed schemes in the partial CSI scenario. However, this increment causes a significant reduction in the weak user performance, as shown in Fig. 4, since in the SUS-SIR [18], the power allocation objective is to maximize the total throughput only. Finally, we can see in Fig. 4 that the PF-MPECG-SIR scheme gives the best weak user performance in both full and partial feedback channel conditions.

The user fairness is tested in Fig. 5 using Jain's fairness index as a function of the number of users K for different NOMA-ZFBF and TDMA-ZFBF systems. We can see that the fairness is decreased for all the schemes as the number of users increased, since increasing the number of users K decreases the opportunity of selecting and serving each user by the BS, and hence, the fairness is decreased among users. In addition, we notice that the fairness for all NOMA-ZFBF schemes is approximately double of that achieved by using the same schemes in TDMA-ZFBF system. This enhancement is due to the fact that NOMA-ZFBF can support two users per beam in each time slot, which is double the number of users that can be supported in TDMA-ZFBF system. However, the PF-MPECG-CORR clearly outperforms PF-MPECG-SIR and all other schemes for both full and partial CSI and for both NOMA and TDMA systems. However, Table 1 provides numerical comparisons between our proposed schemes in terms of the achieved throughput-fairness trade-off over the conventional SUS-SIR scheme [12]. The results are based on Fig. 3 and Fig. 5, when the number of users $K = 150$ users. Furthermore, it's worth to mention that by setting the window size t_c equal to the time horizon W_t, the maximum fairness can be achieved, therefore, we set $W_t = t_c = 225$ for all of our simulation results.

In Fig. 6, we test the overall system sum-rate performance with our proposed schemes versus different numbers of feedback bits and compare each scheme with its full CSI performance. The optimal number of feedback bits can be obtained from the intersections points of partial CSI curves with the full CSI straight lines. We notice that the intersections occur at $B = 14$ bits for PF-MPECG-SIR and PF-MPECG-CORR schemes in NOMA system. On the other hand, the intersection occurs at $B = 12$ bits only for these schemes in TDMA system, since it has a lower sum-rate performance compared to NOMA system.

Fig.7-Fig.8 show the sum-rate performance of NOMA-ZFBF system versus the total transmitted SNR ρ for $K = 50, 150$ users and $B = 4$ bits. In Fig. 7, it can be seen that the throughput performance of both full and partial CSI of our proposed schemes achieve better performance compared to the full CSI performance of Pareto-optimality scheme [15]. In addition, it can be noticed that in high SNR

TABLE 1. Throughput-fairness trade-off enhancements of our proposed algorithms over the SUS-SIR [12]. $K = 150$ users and $t_c = 225$.

	PF-MPECG-SIR		PF-MPECG-CORR		
	Throughput (bps/Hz)	*Fairness Index*	*Throughput (bps/Hz)*	*Fairness Index*	
Full CSI	+ 0.003 (0.02%)	+ 0.211 (33.80%)	- 0.165 (1.13%)	+ 0.318 (50.82%)	Compared to SUS-SIR in NOMA system [12]
Partial CSI	+ 0.046 (0.39%)	+ 0.141 (22.59%)	- 0.050 (0.42%)	+ 0.280 (44.90%)	
Full CSI	+ 1.65 (12.83%)	+ 0.52 (167.8%)	+ 1.49 (11.53%)	+ 0.63 (201.89%)	Compared to SUS-SIR in TDMA system [12]
Partial CSI	+ 1.97 (19.86%)	+ 0.45 (145.6%)	+ 1.87 (18.89%)	+ 0.59 (190.30%)	

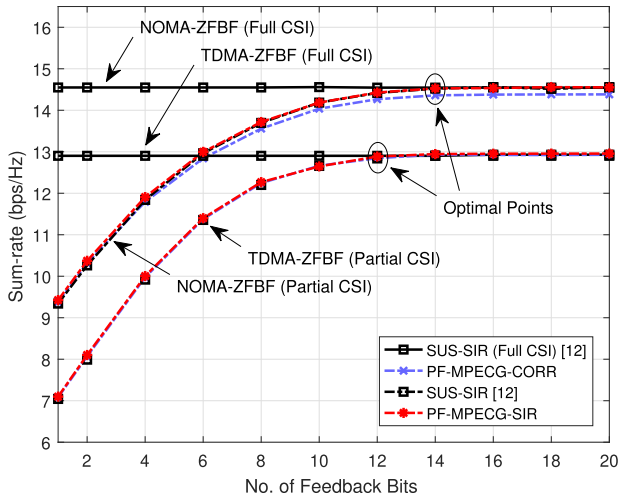

FIGURE 6. Total system sum-rate performance versus different numbers of feedback bits B. The number of users $K = 150$ users.

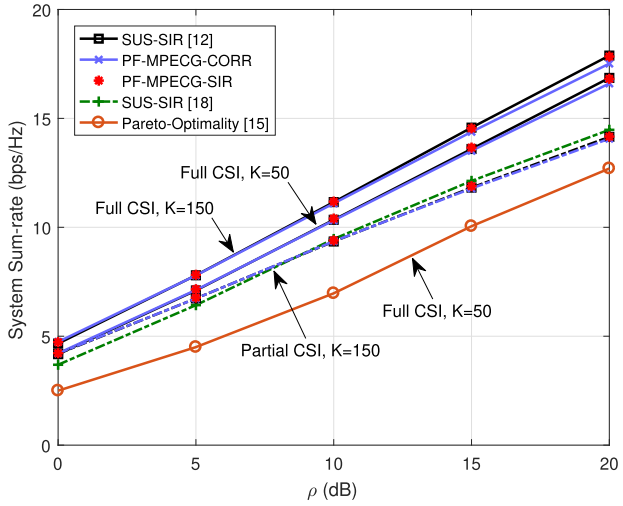

FIGURE 7. Total system sum-rate performance of different NOMA-ZFBF systems versus SNR. $K = 50, 150$ users and $B = 4$ bits/user.

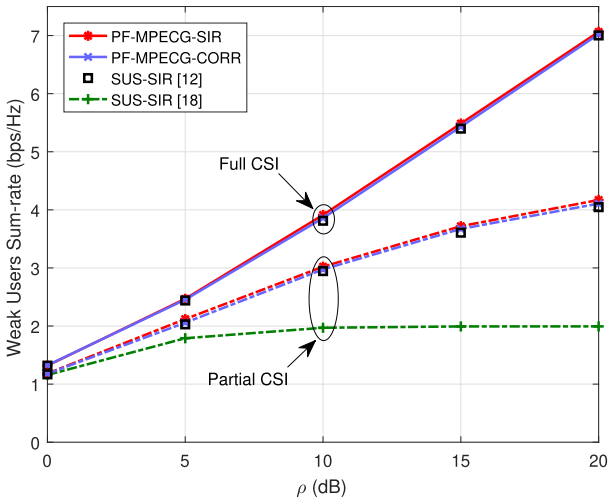

FIGURE 8. Weak users' sum-rate versus SNR for NOMA-ZFBF and TDMA-ZFBF systems. Both systems have identical capacities for all the schemes except the SUS-SIR [18], which represents weak users' sum-rate in NOMA-ZFBF system only. $K = 150$ users and $B = 4$ bits/user.

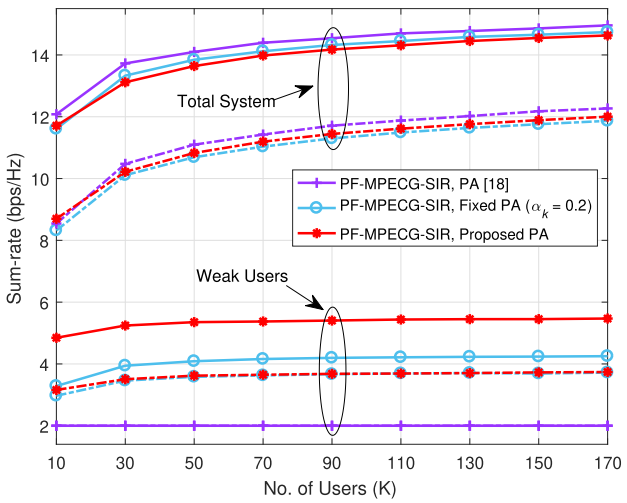

FIGURE 9. Total system and weak users' sum-rates versus the number of users. Different power allocation schemes are applied to the PF-MPECG-SIR. Solid curves refer to the full CSI and the dashed curves refer to the partial CSI with $B = 4$ bits/user.

(more than 10 dB) the performance of SUS-SIR [18], gives slightly higher performance than our proposed schemes in partial CSI for $K = 50$ users. However, the SUS-SIR [18] sacrificing the weak users sum-rate, as shown in Fig. 8, which reflects a similar result to that in Fig. 4.

In Fig. 9, we compare our proposed power PA with two different PA schemes, the proposed PA scheme in [18], and the fixed PA scheme, which allocates a fixed power portion

for each user, i.e., we set $\alpha_k = 0.2$ and $(1-\alpha_k) = 0.8$ for the strong and weak user, respectively. We test the impact of these PA schemes on the sum-rate performance of the total system and the weak users, using our proposed PF-MPECG-SIR as a baseline. In case of full CSI, we notice that the use of the PA [18] and the fixed PA schemes give slightly higher total throughput than that achieved by our proposed PA. On the other hand, the weak users' sum-rate is much higher with our proposed PA compared to those achieved with other PA schemes. In the case of partial CSI, we can observe that the total throughput with our proposed PA became higher than that achieved using the fixed PA. This improvement comes with maintaining the same weak users' throughput obtained using the fixed PA scheme. From simulations, we observe that our proposed PA can guarantee a balance between maximizing the total throughput and maintaining a high sum-rate for the weak users' under both full and partial CSI.

X. CONCLUSION

This paper considered the problem of ZF beam design and user clustering for downlink multiuser NOMA from user fairness perspective under full and partial channel knowledge at the transmitter. We proposed two clustering algorithms based on proportional fairness, PF-MPECG-SIR, and PF-MPECG-CORR. The PF-MPECG-SIR has the ability to maintain the total system throughput with moderate fairness enhancement, while the PF-MPECG-CORR can maximize user fairness with a slight degradation in the total data rate. Furthermore, we proposed an optimal power allocation scheme which can balance between achieving a high system information rate and guaranteeing the QoS of weak users under full and partial feedback channel conditions. Finally, numerical results verified that our proposed algorithms can significantly enhance the throughput-fairness trade-off among users without increasing the computational complexity.

REFERENCES

[1] I. Chih-Lin, C. Rowell, S. Han, Z. Xu, G. Li, and Z. Pan, "Toward green and soft: A 5G perspective," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 66–73, Feb. 2014.
[2] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.
[3] A. Benjebbour, Y. Saito, Y. Kishiyama, A. Li, A. Harada, and T. Nakamura, "Concept and practical considerations of non-orthogonal multiple access (NOMA) for future radio access," in *Proc. ISPACS*, Nov. 2013, pp. 770–774.
[4] K. Higuchi and A. Benjebbour, "Non-orthogonal multiple access (NOMA) with successive interference cancellation for future radio access," *IEICE Trans. Commun.*, vol. 98, no. 3, pp. 403–414, 2015.
[5] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 65, no. 3, pp. 1077–1091, Mar. 2017.
[6] P. Parida and S. S. Das, "Power allocation in OFDM based NOMA systems: A DC programming approach," in *Proc. Globecom Workshops (GC Wkshps)*, Dec. 2014, pp. 1026–1031.
[7] M.-R. Hojeij, J. Farah, C. A. Nour, and C. Douillard, "Resource allocation in downlink non-orthogonal multiple access (NOMA) for future radio access," in *Proc. IEEE Veh. Technol. Conf.*, Glasgow, U.K., May 2015, pp. 1–6.
[8] F. Liu, P. Mahonen, and M. Petrova, "Proportional fairness-based user pairing and power allocation for non-orthogonal multiple access," in *Proc. IEEE Pers., Indoor Mobile Radio Commun. Symp.*, Aug. 2015, pp. 1127–1131.
[9] J. Choi, "Power allocation for max-sum rate and max-min rate proportional fairness in NOMA," *IEEE Commun. Lett.*, vol. 20, no. 10, pp. 2055–2058, Oct. 2016.
[10] M.-R. Hojeij, C. A. Nour, J. Farah, and C. Douillard, "Waterfilling-based proportional fairness scheduler for downlink non-orthogonal multiple access," *IEEE Wireless Commun. Lett.*, vol. 6, no. 2, pp. 230–233, Apr. 2017.
[11] B. Kim, S. Lim, and H. Kim, "Non-orthogonal multiple access in a downlink multiuser beamforming system," in *Proc. IEEE Military Commun. Conf.*, San Diego, CA, USA, Nov. 2013, pp. 18–20.
[12] S. Liu, C. Zhang, and G. Lyu, "User selection and power schedule for downlink non-orthogonal multiple access (NOMA) system," in *Proc. IEEE Int. Conf. Commun. Workshop (ICCW)*, Jun. 2015, pp. 2561–2565.
[13] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE J. Sel. Areas Commun*, vol. 24, no. 3, pp. 528–541, Mar. 2006.
[14] M. Al-Wani, A. Sali, B. M. Ali, A. A. Salah, K. Navaie, C. Y. Leow, N. K. Noordin, and S. J. Hashim, "On short term fairness and throughput of user clustering for downlink non-orthogonal multiple access system," in *Proc. IEEE 89th Veh. Technol. Conf.*, Apr. 2019, pp. 1–6.
[15] J. Seo and Y. Sung, "Beam design and user scheduling for nonorthogonal multiple access with multiple antennas based on Pareto optimality," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2876–2891, Jun. 2018.
[16] Q. Yang, H.-M. Wang, D. W. K. Ng, and M. H. Lee, "NOMA in downlink SDMA with limited feedback: Performance analysis and optimization," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2281–2294, Oct. 2017.
[17] F. Alavi, K. Cumanan, Z. Ding, and A. G. Burr, "Beamforming techniques for nonorthogonal multiple access in 5G cellular networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9474–9487, Oct. 2018.
[18] S. Liu and C. Zhang, "Non-orthogonal multiple access in a downlink multiuser beamforming system with limited CSI feedback," *EURASIP J. Wireless Commun. Netw.*, vol. 2016, no. 1, p. 236, 2016.
[19] Y. Shi, Q. Yu, W. Meng, and Z. Zhang, "Maximum product of effective channel gains: An innovative user selection algorithm for downlink multi-user multiple input and multiple output," *Wireless Commun. Mobile Comput.*, vol. 14, no. 18, pp. 1732–1740, Dec. 2014.
[20] T. Yoo, N. Jindal, and A. Goldsmith, "Multi-antenna downlink channels with limited feedback and user selection," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 7, pp. 1478–1491, Sep. 2007.
[21] N. Jindal, "MIMO broadcast channels with finite-rate feedback," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 5045–5060, Nov. 2006.
[22] R. Jain, D. M. Chiu, and W. R. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer system," Digit. Equip. Corp., Maynard, MA, USA, Tech. Rep. TR-301, 1984.
[23] M. Nasimi, F. Hashim, A. Sali, and R. K. Sahbudin, "QoE-driven cross-layer downlink scheduling for heterogeneous traffics over 4G networks," *Wireless Pers. Commun.*, vol. 96, no. 3, pp. 4755–4780, Oct. 2017.
[24] Y. Xin, Y. H. Nam, Y. Li, and J. C. Zhang, "Reduced complexity precoding and scheduling algorithms for full-dimension MIMO systems," in *Proc. IEEE Global Commun. Conf.*, Dec. 2014, pp. 4858–4863.

MOHANAD M. AL-WANI received the B.S. degree in electrical engineering from Al-Mustansiriya University, Baghdad, Iraq, in 2006, and the M.S. degree in wireless communication engineering from the University of Technology, Baghdad, Iraq, in 2010. He is currently pursuing the Ph.D. degree with Universiti Putra Malaysia. From 2010 to 2011, he was a Lecturer with the Department of Computer Science, Dijlah University College, Baghdad, Iraq. His research interests include new multiple access methods, beamforming, multiuser communication systems, partial feedback techniques, and signal processing for next wireless communication.

ADUWATI SALI received the B.Eng. degree in electrical electronics engineering (communications) from the University of Edinburgh, U.K., in 1999, the M.Sc. degree in communications and network engineering from Universiti Putra Malaysia (UPM), Malaysia, in April 2002, and the Ph.D. degree in mobile and satellite communications from the University of Surrey, U.K., in July 2009. She was the Deputy Director with UPM Research Management Centre (RMC) responsible for research planning and knowledge management from 2016 to 2019. She has been a Professor with the Department of Computer and Communication Systems, Faculty of Engineering, UPM, since February 2019. She was an Assistant Manager with Telekom Malaysia Bhd from 1999 to 2000. She gave consultations to the Malaysian Ministry of Information and Multimedia, the Malaysian Ministry of Higher Education, the National Space Agency (ANGKASA), ATSB Bhd, and Petronas Bhd., on projects related to mobile and satellite communications. In 2014, the fateful event of missing MH370 has requested her to be in printed and broadcasting media, specifically Astro Awani, RTM, TV Al-Hijrah, BERNAMA, Harian Metro, and Metro Ahad, regarding analysis on satellite communication in tracking the aircraft. Her research interests include radio resource management, MAC layer protocols, satellite communications, satellite-assisted emergency communications, IoT systems for environmental monitoring, and 3D video transmission over wireless networks.

She was a recipient of the 2018 Top Research Scientists Malaysia (TRSM) Award from Academy of Sciences Malaysia (ASM). She was involved with EU-IST Satellite Network of Excellence (Sat- NEx) I & II from 2004 to 2009. She is involved with IEEE as a Chair to ComSoc/VTS Malaysia in 2017 and 2018 and a Young Professionals (YP) in 2015; a Young Scientists Network-Academy of Sciences Malaysia (YSN-ASM) as a Chair in 2018 and a Co-Chair for Science Policy in 2017. She is the Principle Investigator and a Collaborator for projects under the local and international funding bodies; namely Malaysian Ministry of Science, Technology and Innovation (MOSTI), Malaysian Ministry of Higher Education (MoHE), Malaysian Communications and Multimedia Commission (MCMC), Research University Grant Scheme (RUGS) (now known as Putra Initiative Grant) UPM, The Academy of Sciences for the Developing World (TWAS-COMSTECH) Joint Grants, EU Horizon2020 Research and Innovation Staff Exchange (H2020-RISE), and NICT Japan - ASEAN IVO. She is also a Chartered Engineer (C.Eng) registered under the U.K. Engineering Council and a Professional Engineer (P.Eng.) under the Board of Engineers Malaysia (BEM).

NOR K. NOORDIN graduated from the University of Alabama, in 1987, and received the Ph.D. degree in wireless and communication engineering from Universiti Putra Malaysia, where she is currently the Dean of engineering. She has published more than 300 journals, book chapters and conference papers. She has led more than 20 research projects, funded by local and international grant providers. Apart from wireless her research interest also includes education engineering.

SHAIFUL J. HASHIM received the B.Eng. degree from the University of Birmingham, U.K., in 1998, the M.Sc. degree from the National University of Malaysia, in 2003, and the Ph.D. degree from Cardiff University, U.K., in 2011, all in electrical and electronics engineering. He is currently an Associate Professor with the Department of Computer and Communication Systems Engineering, Faculty of Engineering, Universiti Putra Malaysia (UPM). He has contributed to more than 100 technical and research publications. His research interests include cloud computing, the Internet of Things (IoT), network security, and non-linear wireless measurement systems. He is one of the winners of the prestigious IEEE MTT-11 2008 Creativity and Originality in Microwave Measurements Competition.

CHEE YEN (BRUCE) LEOW received the B.Eng. degree in computer engineering from Universiti Teknologi Malaysia (UTM), in 2007, and the Ph.D. degree from Imperial College London, in 2011. Since July 2007, he has been an Academic Staff with the School of Electrical Engineering, Faculty of Engineering, Universiti Teknologi Malaysia (UTM). He is currently an Associate Professor with the Faculty and a Research Fellow in the Wireless Communication Centre (WCC), Higher Institution Centre of Excellence, UTM, and also with the UTMEricsson Innovation Centre for 5G. He is also the Secretary of Malaysia IMT and the Future Networks Working Group and also the Head of the Technology Subgroup in Malaysia 5G Task Force. He is among the pioneers of 5G initiatives in Malaysia. His research interests include non-orthogonal multiple access, cooperative communication, UAV communication, MIMO, hybrid beamforming, physical layer security, wireless power transfer, and prototype development using software defined radio, for 5G and the IoT applications.

IOANNIS KRIKIDIS (S'03–M'07–SM'12–F'19) received the Diploma degree in computer engineering from the Computer Engineering and Informatics Department (CEID), University of Patras, Greece, in 2000, and the M.Sc. and Ph.D. degrees from the Ecole Nationale Superieure des Telecommunications (ENST), Paris, France, in 2001 and 2005, respectively, all in electrical engineering. From 2006 to 2007, he was a Postdoctoral Researcher with ENST, Paris, France, and from 2007 to 2010, he was a Research Fellow with the School of Engineering and Electronics, University of Edinburgh, Edinburgh, U.K. He has held also research positions at the Department of Electrical Engineering, University of Notre Dame; the Department of Electrical and Computer Engineering, University of Maryland; the Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg; and the Department of Electrical and Electronic Engineering, Niigata University, Japan. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, University of Cyprus, Nicosia, Cyprus. His current research interests include wireless communications, cooperative networks, 4G/5G communication systems, wireless powered communications, and secrecy communications. He was a recipient of the Young Researcher Award from the Research Promotion Foundation, Cyprus, in 2013, and the IEEE ComSoc Best Young Professional Award in Academia, in 2016. He serves as an Associate Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING, and the IEEE WIRELESS COMMUNICATIONS LETTERS. He has been recognized by Thomson Reuters as an ISI Highly Cited Researcher 2017 & 2018.