Matrix Normalization Based ZF Hybrid Precoded Multi-User MIMO mmWave Systems with Massive Array

Hang Li, Thomas Q. Wang, Xiaojing Huang, and Y. Jay Guo
Global Big Data Technologies Center (GBDTC), University of Technology, Sydney, Australia
Emails: {Hang.Li, Qian.Wang, Xiaojing.Huang, Jay.Guo}@uts.edu.au

***Abstract*—The superiority of exploring millimeter wave (mmWave) frequencies for future wireless communication systems has pushed forward the development of large-scale antenna arrays for achieving sufficient array gain and high spectral efficiency. In this paper, we study the matrix normalization (MN) based zero-forcing (ZF) hybrid precoding in multi-user multi-input-multi-output (MU-MIMO) mmWave systems. We derive the upper bounds of the achievable rate for two representative hybrid array structures, i.e., fully-connected structure and partially-connected structure. Analytical and simulation results validate the tightness of the proposed performance upper bounds for both hybrid structures using massive array, and provide a comparison of the achievable rate using MN and vector normalization (VN).**

***Index Terms*—Matrix normalization, vector normalization, ZF hybrid precoding, mmWave massive MIMO.**

I. Introduction

To balance the system performance and hardware cost, millimeter wave (mmWave) beamforming with massive hybrid antenna array [1], [2] has been regarded as an attractive solution for 5G wireless communication systems. In view of the mapping from antenna elements to radio frequency (RF) chains, hybrid array architecture can be classified into the fully-connected structure [1] where each antenna connects to multiple phase shifters and all RF chains, and the partially-connected structure [2] where each antenna only connects to one phase shifter and one RF chain. The fully-connected structure employs full beamforming gain for each RF chain such that it can approach the performance of a fully digital scheme with much fewer number of RF chains [3]. On the other hand, the partially-connected structure is more energy-efficient [4], and hence preferable for practical deployment with massive antennas at the cost of some performance loss.

Compared with the nonlinear precoding, e.g., optimal dirty paper coding (DPC) [5] that is implemented with significant additional complexity, linear precoding schemes such as zero-forcing (ZF) [6] are considered as simple and near-optimal methods in massive MIMO systems for multiuser interference cancellation when the channel is available. For ZF precoding, there are two power normalization methods, i.e., matrix normalization (MN) and vector normalization (VN), commonly used in the practical case of power allocation among different data streams. *Matrix normalization* regulates the precoding matrix by multiplying a scalar such that the power constraint at the base station (BS) is satisfied, which results in equal receive power for each user. Therefore, it provides better trade-off in fairness. Alternatively, equal transmit power can be imposed across all data streams by normalizing the precoding matrix with different scalars while meeting the power constraint, known as *vector normalization*. The corresponding sum rate performance was analysed and compared under Rayleigh fading channels in [7], [8].

ZF hybrid precoding has been applied in MU-MIMO mmWave systems [9], [10], which requires a small amount of training and feedback overhead to obtain the equivalent baseband channel state information. The results in [10] showed that the performance of ZF hybrid precoding based on VN approached that of the unconstrained digital block diagonalization precoding with relatively small codebooks. However, the performance analysis was only given in terms of the fully-connected structure without any comparison with the partially-connected structure. Also, the performance difference between MN and VN for ZF hybrid precoding in MU-MIMO mmWave systems with massive array has not been studied yet.

In this paper, with regard to fully-connected structure and partially-connected structure, we derive the upper bounds of achievable rate of matrix normalization based ZF hybrid precoding in typical mmWave channels. Analytical and simulation results show that the proposed performance upper bounds are tight for both structures, particularly when the number of array antennas is in the large dimensional regime. Additionally, we present an analytical comparison between MN and VN, which shows that MN provides a notion of fairness at a negligible rate loss compared with VN.

Notations: $\mathbf{A}$, $\mathbf{a}$ and a stand for a matrix, a column vector and a scalar, respectively; $\mathbf{A}_{i,j}$ is the entry on the ith row and jth column of $\mathbf{A}$; $\mathbf{A}^T$ and $\mathbf{A}^H$ denote the transpose and conjugate transpose of $\mathbf{A}$, respectively. $\|\mathbf{A}\|_F$ is the Frobenius norm of $\mathbf{A}$, and $\mathbf{I}_N$ is the identity matrix with N dimensions; $\mathcal{N}(\mathbf{m}, \mathbf{V})$ represents a complex Gaussian random vector with mean $\mathbf{m}$ and covariance matrix $\mathbf{V}$. Further, the notations $\log(\cdot)$, $\mathbb{E}[\cdot]$, $\mathrm{Tr}(\cdot)$ and $|(\cdot)|$ represent the logarithmic, expectation, trace and absolute value of $(\cdot)$, respectively.

This research was supported under Australian Research Council's Discovery Projects funding scheme (project number DP160101693).

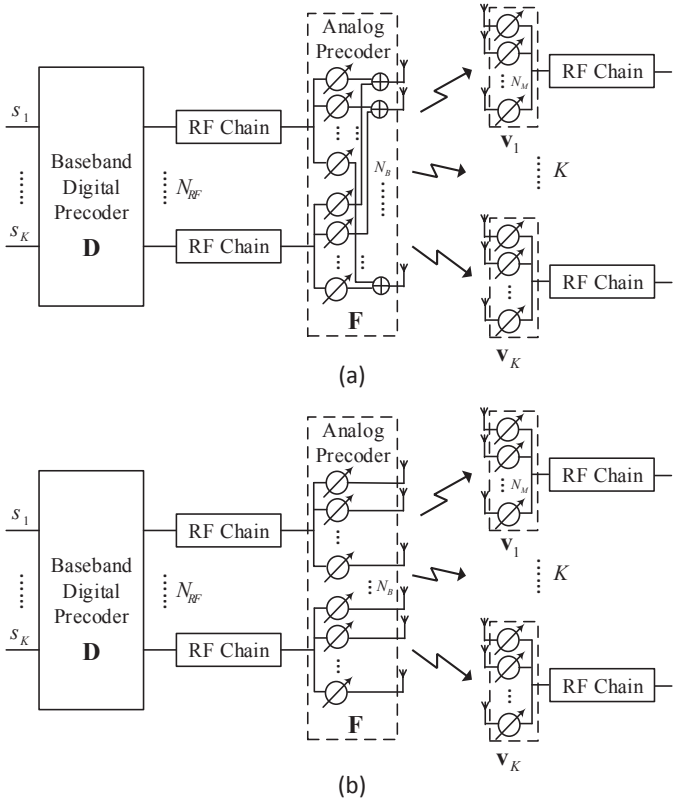

Fig. 1. Two typical structures of the hybrid precoding: (a) Fully-connected structure, where each antenna connects to multiple phase shifters and all RF chains. (b) Partially-connected structure, where each antenna only connects to one phase shifter and one RF chain.

II. System Model

As in [10], we consider a narrowband multiuser downlink transmission system, which consists of a BS and K users. The BS is equipped with N_{RF} RF chains and N_B antenna elements ($K \leq N_{RF} \ll N_B$), and each user is equipped with N_M antennas. We assume that the BS processes only one data stream with each user (i.e., the total number of data streams equals K). For simplicity, we also assume that $N_{RF} = K$, and only one RF chain is used at each user side due to the limited processing capacity. As shown in Fig. 1, we consider both fully-connected and partially-connected structures, which use different analog precoding structures thus leading to different achievable rates.

We denote the transmitted symbols of K users as $\mathbf{s} = [s_1, s_2, ..., s_K]^T$, where $\mathbf{s} \sim \mathcal{N}(\mathbf{0}, \frac{P_t}{K}\mathbf{I}_K)$ and P_t is the average total transmit power. Let $\mathbf{H}_k$ denote the $N_M \times N_B$ mmWave channel matrix between the BS and user k, and $\mathbf{n}_k \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_{N_M})$ denote the additive white Gaussian noise of user k's received signal. Therefore, the received signal vector of user k after analog beamforming combination can be expressed as

$$\mathbf{y}_k = \mathbf{v}_k^H \mathbf{H}_k \mathbf{F}\mathbf{D}\mathbf{s} + \mathbf{v}_k^H \mathbf{n}_k, \quad (1)$$

where $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_K]$ is the $K \times K$ baseband digital precoding matrix for multiuser interference cancellation, and $\mathbf{F}$ is the $N_B \times K$ RF analog precoding matrix. For the fully-connected structure, $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_K]$, each entry with normalized constant modulus $\sqrt{1/N_B}$ for implementing analog phase shifters. For the partially-connected structure, $\mathbf{F}$ is block diagonal, and expressed as

$$\mathbf{F} = \begin{pmatrix} \mathbf{f}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{f}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{f}_K \end{pmatrix},$$

where $\mathbf{f}_K$ is an $\frac{N_B}{K} \times 1$ vector with $|\mathbf{F}_{k,n}| = \sqrt{K/N_B}$, and $\frac{N_B}{K}$ is assumed to be an integer. To satisfy the total power constraint, we have $\|\mathbf{F}\mathbf{D}\|_F^2 = K$. $\mathbf{v}_k$ is the $N_M \times 1$ receive analog beamforming vector, each entry with constant modulus $\sqrt{1/N_M}$.

It is assumed that the channel $\mathbf{H}_k$ can be decomposed into a deterministic channel matrix $\mathbf{H}_k^L$ induced by line-of-sight (LOS) and a random channel matrix $\mathbf{H}_k^N$ induced by scattering components [9], [11],

$$\begin{aligned} \mathbf{H}_k =& \mathbf{H}_k^L + \mathbf{H}_k^N \\ =& \sqrt{\frac{N_B N_M \eta_k}{1+\eta_k}} e^{j\beta_k} \mathbf{a}_M(\theta_k^L)\mathbf{a}_B^H(\phi_k^L) \\ &+ \sqrt{\frac{N_B N_M}{N_c N_{sc}(1+\eta_k)}} \sum_{c=1}^{N_c}\sum_{p=1}^{N_{sc}} \alpha_k^{c,p} \mathbf{a}_M(\theta_k^{c,p})\mathbf{a}_B^H(\phi_k^{c,p}), \end{aligned} \quad (2)$$

where η_k is known as the Rician factor of user k and $\beta_k \in [0, 2\pi]$ is the random phase. The scattering component $\mathbf{H}_k^N$ consists of N_c scattering clusters and N_{sc} scatters within each cluster. $\alpha_k^{c,p} \sim \mathcal{N}(0,1)$, $\theta_k^{c,p}$ and $\phi_k^{c,p}$ are the complex gain, angles of arrival and departure (AoAs/AoDs) for the pth path in the cth cluster, respectively. Further, $\mathbf{a}_M(\theta_k^{c,p})$ and $\mathbf{a}_B^H(\phi_k^{c,p})$ are the corresponding normalized antenna array response vectors of the BS and user k, respectively. For an uniform linear array (ULA), we have

$$\mathbf{a}_B(\phi) = \frac{1}{\sqrt{N_B}}[1, e^{j\frac{2\pi}{\lambda}d\sin(\phi)}, ..., e^{j\frac{2\pi}{\lambda}(N_B-1)d\sin(\phi)}]^T, \quad (3)$$

where λ is the carrier wavelength, and d is the adjacent element spacing. $\mathbf{a}_M(\theta)$ can be written in a similar fashion. The analysis in this paper can be also applied to an uniform planar array.

Mmwave channel model is characterized by the LOS component and non-negligible scattering components, where LOS path dominates the power distribution across the multipath. Especially when the Rician factor is large, it can be assumed that the AOA of LOS signal is approximately regarded as the AOA of received signals, which greatly simplifies signal processing. On the other hand, $\alpha_k^{c,p}$, $\theta_k^{c,p}$ and $\phi_k^{c,p}$ are more randomly distributed [9], [11] between the BS and the users as the number of scattering clusters increases. Accordingly, we consider that the entries of random channel matrix $\mathbf{H}_k^N$ are approximately independent and identically distributed (i.i.d.) random variables $\sim \mathcal{N}(0, \frac{1}{1+\eta_k})$.

III. Hybrid Precoding

A. RF Beamforming

Since phase shifters are digitally implemented, RF beamforming angles may be chosen from finite-size codebook [10]. Specifically, in this paper, we use the beamsteering codebooks for RF beamforming design, which have the same form as antenna array response vectors in (3), to simplify the codebook design due to single parameter quantization. Here, we also consider that one RF chain precoding is designed for one user only (i.e., each beamforming signal from one RF chain potentially points at each user.) as in [10], thus maximizing the desired signal power of each user, and ignoring the mutual interference among users. Therefore, the optimal transmit and receive RF beamforming vectors for user k, $\mathbf{b}_B(\phi_k^*)$ and $\mathbf{b}_M(\theta_k^*)$, can be selected from $\mathcal{W}$ and $\mathcal{V}$ which are the beamsteering codebooks of the BS and users respectively, such that

$$\{\mathbf{b}_B(\phi_k^*), \mathbf{b}_M(\theta_k^*)\} = \underset{\substack{\forall \mathbf{b}_B(\phi_k)\in\mathcal{W} \\ \forall \mathbf{b}_M(\theta_k)\in\mathcal{V}}}{\operatorname{argmax}} |\mathbf{b}_M^H(\theta_k)\mathbf{H}_k\mathbf{b}_B(\phi_k)|. \quad (4)$$

It is noted that the selection process can be implemented by searching the codebooks with efficient beam training algorithms developed in [12], such that the strongest path AOA/AOD can be obtained. As shown in [13], the optimal singular-value-decomposition (SVD) transmit and receive RF beamforming vectors for the channel in (2) converge to the array response vectors in the strongest direction with massive array. Hence, the searching process can be regarded as near-optimal searching to find the LOS direction. Let $\mathbf{v}_k = \mathbf{b}_M(\theta_k^*)$, $\mathbf{f}_k = \mathbf{b}_B(\phi_k^*)$ for the fully-connected structure, and $\mathbf{f}_k = \sqrt{K}\hat{\mathbf{b}}_B(\phi_k^*)$ for the partially-connected structure, where $\hat{\mathbf{b}}_B(\phi_k^*)$ is composed of the first $\frac{N_B}{K}$ entries of $\mathbf{b}_B(\phi_k^*)$.

B. Baseband ZF Precoding

We define that $\mathbf{H}_{eq} = [\mathbf{h}_{eq,1}, \mathbf{h}_{eq,2}, ..., \mathbf{h}_{eq,K}]^H$ is the equivalent baseband channel matrix between the BS's RF chains and all users' RF chains, where $\mathbf{h}_{eq,k}^H = \mathbf{v}_k^H\mathbf{H}_k\mathbf{F}$. $\mathbf{H}_{eq}$ can be directly estimated by exploiting the channel reciprocity [9], or estimated by the users and then fed back to the BS [10]. Since the number of RF chains, N_{RF}, is much smaller than that of the transmit antennas N_B, the required feedback overhead and inverse matrix calculation complexity can be greatly reduced. Assuming that $\mathbf{H}_{eq}$ is perfectly estimated in the high SNR regime, the non-normalized digital ZF precoding matrix $\mathbf{D}_1$ is given by

$$\mathbf{D}_1 = \mathbf{H}_{eq}^H(\mathbf{H}_{eq}\mathbf{H}_{eq}^H)^{-1}. \quad (5)$$

To satisfy the power constraint, we have two normalization methods (i.e., matrix/vector normalizations) to normalize $\mathbf{D}_1$. To guarantee each user with equal receive power, MN is to multiply $\mathbf{d}_{1,k}$ by $\sqrt{K}\|\mathbf{F}\mathbf{D}_1\|_F^{-1}$, whereas VN is to multiply $\mathbf{d}_{1,k}$ by $\|\mathbf{F}\mathbf{d}_{1,k}\|_F^{-1}$ to keep equal transmit power for each user, where $\mathbf{d}_{1,k}$ denotes the kth column of $\mathbf{D}_1$. Although they have different normalization factors, multi-user interference can be totally cancelled due to $\mathbf{h}_{eq,i}^H\mathbf{d}_j = 0, \forall i \neq j$.

C. Asymptotic Downlink Achievable Rate

In the following, we derive the achievable rate upper bounds according to MN assuming perfect equivalent baseband channel. For simplicity, we assume that all the users have the same Rician factor, i.e., $\eta_k = \eta, \forall k$. Using the ZF hybrid precoding described in Section III, the average achievable rate per user, R^{MN}, is given by

$$R^{MN} = \mathbb{E}\left[\log\left(1 + \frac{P_t}{\sigma^2\|\mathbf{F}\mathbf{D}_1\|_F^2}\right)\right], \quad (6)$$

which is upper bounded by

1) For the fully-connected structure:

$$R_f^{MN} \leq R_f^{up} = \log_2\left[1 + \frac{P_t}{\sigma^2}\cdot\frac{N_BN_M\eta + K}{K(1+\eta)}\right], \quad (7)$$

and 2) For the partially-connected structure:

$$R_p^{MN} \leq R_p^{up} = \log_2\left[1 + \frac{P_t}{\sigma^2}\cdot\frac{N_BN_M\eta\|\mathbf{F}_f^H\mathbf{F}_p\|_F^2 + K^2}{K^2(1+\eta)}\right]. \quad (8)$$

The proof of (7) and (8) is provided in the Appendix. Eq. (7) shows that, for the fully-connected structure, the upper bound is irrelevant to analog precoding matrix $\mathbf{F}$, but for the partially-connected structure in (8), it depends on $\mathbf{F}$ designed in Section III.A. On the other hand, the upper bound depends on the Rician factor, and it is a monotonically increasing function in term of η provided that $N_BN_M \geq K$. Particularly, when $\eta \to +\infty$ (i.e., single-path channels), the upper bounds are approximated by $\log_2\left[1 + \frac{P_t}{\sigma^2}\cdot\frac{N_BN_M}{K}\right]$ and $\log_2\left[1 + \frac{P_t}{\sigma^2}\cdot\frac{N_BN_M\|\mathbf{F}_f^H\mathbf{F}_p\|_F^2}{K^2}\right]$, respectively. The proposed upper bounds provide insights into the performance of the MN based ZF hybrid precoding in MU-MIMO mmWave systems which will be validated by the simulations in Section IV.

We are interested in the massive array as the transceiver generally requires large antenna arrays to achieve received power gain for typical mmWave channels. Therefore, it is meaningful to evaluate the asymptotic upper bounds for the case with a large number of antennas. When the number of transmit antennas $N_B \to \infty$, we have $\|\mathbf{F}_f^H\mathbf{F}_p\|_F^2 \to 1$ since $\mathbf{F}_f^H\mathbf{F}_p \to \frac{1}{\sqrt{K}}\mathbf{I}_K$, where $\mathbf{F}_f$ and $\mathbf{F}_p$ are analog precoding matrices for two structures, respectively. It is worth noting that the diagonal elements of $\mathbf{F}_f^H\mathbf{F}_p$ are exactly $\frac{1}{\sqrt{K}}$ while the off-diagonal elements can be approximated as a summation of $\frac{N_B}{K}$ independent unit-norm complex numbers, which suggests that the norm of off-diagonal elements is much less than $\frac{1}{\sqrt{K}}$ with very high probability when $N_B \to \infty$. As a result, $\mathbf{F}_f^H\mathbf{F}_p \to \frac{1}{\sqrt{K}}\mathbf{I}_K$. Therefore, the asymptotic achievable rate per user for the partially-connected structure in (8) is approximately bounded by $\log_2\left[1 + \frac{P_t}{\sigma^2}\cdot\frac{N_BN_M\eta+K^2}{K^2(1+\eta)}\right]$, and the upper bounds gap, ΔR^{up}, between two structures satisfies

$$\Delta R^{up} = R_f^{up} - R_p^{up} \overset{(a)}{\lesssim} \log_2\left(\frac{N_BN_M\eta K + K^2}{N_BN_M\eta + K^2}\right) \overset{(b)}{\approx} \log_2 K, \quad (9)$$

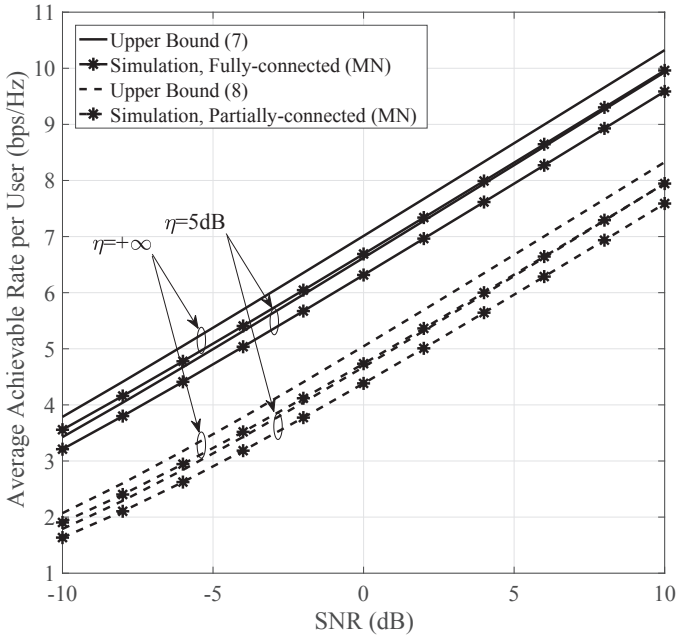

Fig. 2. Average achievable rate per user versus the SNR, where $N_B = 128$, $N_M = 4$ and $K = 4$.

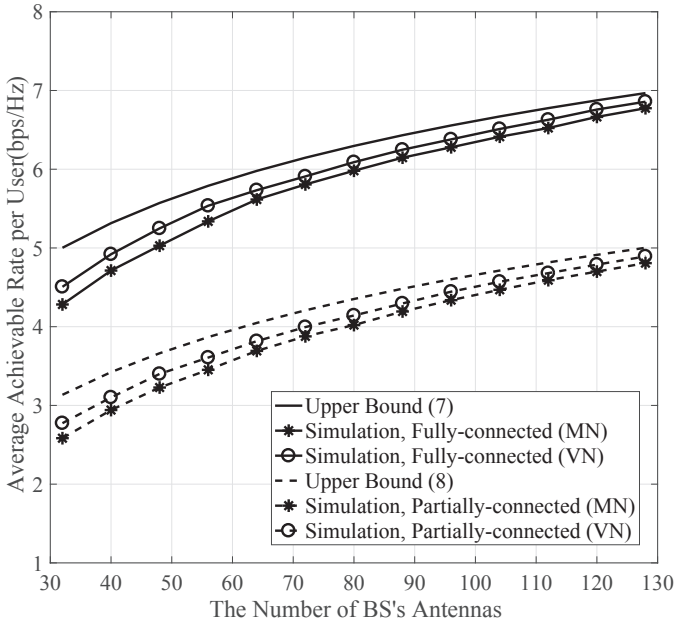

Fig. 3. Average achievable rate per user versus the number of BS's antennas, where $N_M = 4$, $K = 4$ and $\eta = 15$dB.

where (a) is derived by noting that any positive numbers a and b with $a \geq b$, satisfy $\log_2\left(\frac{1+a}{1+b}\right) \leq \log_2\left(\frac{a}{b}\right)$. (b) can be obtained assuming $N_B N_M \eta \gg K^2$.

Note that our proposed achievable rate upper bound is tighter than the one proposed in [9, eq. (23)] since $\|\mathbf{F}_f^H \mathbf{F}_f\|_F^2 \geq K$, which can be derived from the fact that if $\mathbf{A}$ and $\mathbf{B}$ are real positive semi-definite matrices of the same size then $\mathrm{Tr}(\mathbf{A}^2)\,\mathrm{Tr}(\mathbf{B}^2) \geq [\mathrm{Tr}(\mathbf{AB})]^2$ and letting $\mathbf{A} = \mathbf{F}_f^H \mathbf{F}_f$ and $\mathbf{B} = \mathbf{I}_K$.

IV. Analytical and Simulation Results

The proposed schemes are simulated using a hybrid ULA with $d = \lambda/2$. We assume that each user experiences multipath fading channel in cluster, where the number of multipath is 7 with one LOS path and the remaining scatters, and the equivalent baseband channel is perfectly estimated. The AoA/AOD of any user's signal is assumed to be uniformly distributed over $[0, 2\pi]$.

Fig. 2 shows the average achievable rates and the corresponding upper bounds versus the SNR, which is defined as SNR = $\frac{P_t}{\sigma^2}$, for two structures respectively. It is observed that the fully-connected structure performs better than partially-connected structure, and their average achievable rates are close to the proposed corresponding upper bounds in the large numbers of antennas regime. The achievable rate gap between two structures approaches $\log_2 K$ = 2bps/Hz with the increase of SNR as expected in (9). With a sufficiently large Rician factor, the higher average achievable rate can be obtained, as the transceiver exploits more gain from the LOS component and the interference from other users is further suppressed.

Fig. 3 compares the average achievable rates for MN/VN versus the number of BS's antennas given SNR = 0 dB. It

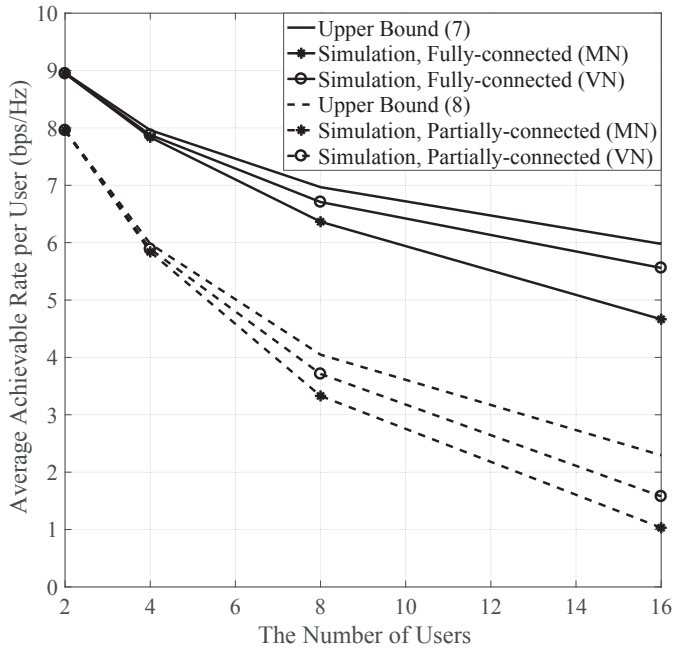

Fig. 4. Average achievable rate per user versus the number of users, where $N_B = 256$, $N_M = 4$ and $\eta = 15$dB.

can be seen that the gap between the average achievable rate and the corresponding upper bound becomes smaller with the increase of the number of BS's antennas for both hybrid structures, which verifies the tightness of the derived upper bounds in (7) and (8) with massive array. It is also shown that the achievable rate with VN is higher than that of MN. The reason is as follows.

For ZF hybrid precoding, the average achievable rate per

user with VN can be expressed as

$$R^{VN} = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\log_2\left(1+\frac{P_t}{K\sigma^2\|\mathbf{F}\mathbf{d}_{1,k}\|_F^2}\right)\right] \geq \mathbb{E}\left[\log_2\left(1+\frac{P_t}{\sigma^2\sum_{k=1}^{K}\|\mathbf{F}\mathbf{d}_{1,k}\|_F^2}\right)\right] = R^{MN}, \tag{10}$$

where $\sum_{k=1}^{K}\|\mathbf{F}\mathbf{d}_{1,k}\|_F^2 = \|\mathbf{F}\mathbf{D}_1\|_F^2$. (10) can be derived by using the *Arithmetic-geometric Inequality* defined in [14]. The MN based ZF hybrid precoding presents a fairness provisioning precoder in spite of 2.5% average achievable rate loss compared with VN.

Fig. 4 shows the average achievable rates for MN/VN versus the number of users given SNR = 0 dB. As shown in Fig. 4, the achievable rate per user decreases with the increasing number of users, as the power allocated to each user is decreased. It also shows that the performance of VN is better than that of MN, and the gap between the average achievable rate and the corresponding upper bound grows with the number of users. The upper bounds gap between two structures shown in the figure demonstrates the correctness of (9).

V. Conclusion

In this paper, in terms of the fully-connected structure and partially-connected structure, we derive the achievable rate upper bounds using MN based ZF hybrid precoding in typical mmWave channels. Numerical results show the tightness of the proposed performance upper bounds for both hybrid structures using massive array. It is shown although MN has limited performance loss compared with VN, it provides strict fairness for all users.

Appendix

Using *Jensen's Inequality* of concave function, we have

$$\mathbb{E}\left[\log_2\left(1+\frac{P_t}{\sigma^2\|\mathbf{F}\mathbf{D}_1\|_F^2}\right)\right] \leq \log_2\left(1+\frac{P_t}{\sigma^2}\cdot\mathbb{E}\left\{\frac{1}{\|\mathbf{F}\mathbf{D}_1\|_F^2}\right\}\right), \tag{11}$$

where $\log_2(1+x)$ is concave function and x is a random variable. Therefore, the derivation of upper bound is equivalently to obtain $\mathbb{E}\left\{\frac{1}{\|\mathbf{F}\mathbf{D}_1\|_F^2}\right\}$. By substituting (5) into it, we have

$$\begin{aligned}
&\mathbb{E}\left\{\frac{1}{\|\mathbf{F}\mathbf{D}_1\|_F^2}\right\}\\
=&\mathbb{E}\left\{\left[\mathrm{Tr}[(\mathbf{H}_{eq}\mathbf{H}_{eq}^H)^{-H}\mathbf{H}_{eq}\mathbf{F}^H\mathbf{F}\mathbf{H}_{eq}^H(\mathbf{H}_{eq}\mathbf{H}_{eq}^H)^{-1}]\right]^{-1}\right\}\\
\overset{(a)}{\leq}&\frac{1}{K^2}\mathbb{E}\{\mathrm{Tr}[\mathbf{H}_{eq}(\mathbf{F}^H\mathbf{F})^{-1}\mathbf{H}_{eq}^H]\}\\
=&\frac{1}{K^2}\mathbb{E}\left\{\sum_{k=1}^{K}\mathbf{v}_k^H\mathbf{H}_k\mathbf{F}(\mathbf{F}^H\mathbf{F})^{-1}\mathbf{F}^H\mathbf{H}_k^H\mathbf{v}_k\right\}\\
\overset{(b)}{=}&\frac{1}{K^2}\left\{\sum_{k=1}^{K}\mathbf{v}_k^H\mathbf{H}_k^L\mathbf{F}(\mathbf{F}^H\mathbf{F})^{-1}\mathbf{F}^H(\mathbf{H}_k^L)^H\mathbf{v}_k\right.\\
&\left.+\sum_{k=1}^{K}\mathbf{v}_k^H\mathbb{E}[\mathbf{H}_k^N\mathbf{F}(\mathbf{F}^H\mathbf{F})^{-1}\mathbf{F}^H(\mathbf{H}_k^N)^H]\mathbf{v}_k\right\}\\
\overset{(c)}{=}&\frac{1}{K^2}\left\{\frac{N_BN_M\eta}{1+\eta}\sum_{k=1}^{K}\mathbf{a}_B^H(\phi_k^L)\mathbf{F}(\mathbf{F}^H\mathbf{F})^{-1}\mathbf{F}^H\mathbf{a}_B(\phi_k^L)\right.\\
&\left.+\frac{K}{1+\eta}\mathrm{Tr}[\mathbf{F}(\mathbf{F}^H\mathbf{F})^{-1}\mathbf{F}^H]\right\},
\end{aligned} \tag{12}$$

where (a) holds due to the property of an $N\times N$ positive definite matrix $\mathbf{A}$ with $\frac{N}{\mathrm{Tr}(\mathbf{A}^{-1})}\leq\frac{\mathrm{Tr}(\mathbf{A})}{N}$. By substituting (2) into (12), we remove the cross terms between $\mathbf{H}_k^L$ and $\mathbf{H}_k^N$ due to independent of each other and the entries of $\mathbf{H}_k^N\sim\mathcal{N}(0,\frac{1}{1+\eta})$, and thus (b) is derived. We assume that $\phi_k^L\approx\phi_k^*$ and $\theta_k^L\approx\theta_k^*$ as in [9] since the LOS path rules the power allocation among all paths, such that the estimated strongest AOA/AOD is close to that of LOS and thus we have (c). By substituting $\mathbf{F}$ of two structures into (12), we complete the proof of (7) and (8).

References

[1] R. W. Heath, Jr., N. Gonzlez-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, Apr. 2016.

[2] J. Zhang, X. Huang, V. Dyadyuk, and Y. Guo, "Massive hybrid antenna array for millimeter-wave cellular communications," *IEEE Wireless Commun.*, vol. 22, no. 1, pp. 79-87, Feb. 2015.

[3] 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.

[4] X. Gao, L. Dai, S. Han, C.-L. I, and R. W. Heath, Jr., "Energy-efficient hybrid analog and digital precoding for mmWave MIMO systems with large antenna arrays," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 998-1009, Apr. 2016.

[5] M. Costa, "Writing on dirty paper," *IEEE Trans. Inf. Theory*, vol. IT-29, pp. 439-441, May 1983.

[6] 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.

[7] M. Ghosh, "A comparison of normalizations for ZF precoded MU-MIMO systems in multipath fading channels," *IEEE Wireless Commun. Lett.*, vol. 2, no. 5, pp. 515-518, Oct. 2013.

[8] Y. Lim, C. Chae, and G. Caire, "Performance analysis of massive MIMO for cell-boundary users," *IEEE Trans. Wireless Commun.*, vol. 14, no. 12, pp. 6827-6843, Dec. 2015.

[9] L. Zhao, D. W. K. Ng, and J. Yuan, "Multi-user Precoding and Channel Estimation for Hybrid Millimeter Wave Systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1576-1590, Jul. 2017.

[10] A. Alkhateeb, G. Leus, and R. W. Heath, "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6481-6494, Nov. 2015.

[11] S. Buzzi and C. DAndrea, "Doubly massive mmWave MIMO systems: Using very large antenna arrays at both transmitter and receiver," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016.

[12] S. Hur et al., "Millimeter wave beamforming for wireless backhaul and access in small cell networks," *IEEE Trans. Commun.*, vol. 61, no. 10, pp. 4391-4403, Oct. 2013.

[13] O. El Ayach, R. W. Heath, Jr., S. Abu-Surra, S. Rajagopal, and Z. Pi, "The capacity optimality of beam steering in large millimeter wave MIMO systems," in *Proc. 2012 IEEE International Workshop Signal Process. Advances Wireless Commun.*, pp. 100-104.

[14] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed. San Diego, CA, USA: Academic, 2007.