IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 21, 2017, accepted July 10, 2017, date of publication July 13, 2017, date of current version August 22, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2726528

Three-Dimension Massive MIMO for Air-to-Ground Transmission: Location-Assisted Precoding and Impact of AoD Uncertainty

YOUYUN XU[1,2], (Senior Member, IEEE), XIAOCHEN XIA[3], KUI XU[3], (Member, IEEE), AND YURONG WANG[3]

[1]National Engineering Research Center of Communication and Network Technologies, Nanjing University of Posts and Telecommunications, Nanjing 0086-210003, China
[2]Institute of Wireless Communication Technology, Shanghai Jiao Tong University, Shanghai 0086-200240, China
[3]Institute of Communication Engineering, PLA University of Science and Technology, Nanjing 0086-210007, China

Corresponding author: Xiaochen Xia (tjuxiaochen@gmail.com)

This work was supported in part by the Major Research Plan of National Natural Science Foundation of China under Grant 91438115, in part by the National Natural Science Foundation of China under Grant 61671472 and Grant 61371123, and in part by the Jiangsu Province Natural Science Foundation under Grant BK20160079. This work has been presented at the IEEE INFOCOM 2017 Workshop on 5G&Beyond-Enabling Technologies and Applications, Atlanta, GA, USA.

ABSTRACT This paper investigates the 3-D massive multiple-input multiple-output (MIMO) for air-to-ground transmission, where an air platform (AP) is equipped with a 2-D rectangular antenna array and communicates with a number of user equipments (UEs) on the ground. By exploiting the slow time-varying parameters, such as channel correlation and angles of departure (AoDs) of UEs, we first propose a location-assisted two-layer precoding scheme for downlink transmission. The first-layer precoding aims to decompose the original massive MIMO system into several low-dimension MIMO systems, with each operating on the orthogonal subspace. Through proper UE clustering, we show that the first-layer precoding matrix can be approximated using a constant-envelope matrix, which results in significant reduction on hardware complexity of AP. The second-layer precoding is designed to eliminate the multi-UE interference within each low-dimension MIMO system. Since the AoD information is usually not perfectly known at AP, we then investigate the effect of AoD uncertainty on the performance of the proposed precoding scheme. In particular, we propose a new analytic method to fast estimate approximately the power loss due to AoD error. Numerical simulations are presented to evaluate the performance of location-assisted precoding under different system parameters, including Rician factor, altitude of AP, and AoD uncertainty. The results show that the location-assisted precoding outperforms match filter precoding and basis expansion-based precoding in the air-to-ground transmission scenarios significantly.

INDEX TERMS Air-to-ground transmission, three-dimension massive MIMO, location-assisted precoding, AoD uncertainty, spectral efficiency.

I. INTRODUCTION

Seamless wide-area coverage is the basic scenario of mobile communications. In this scenario, one main challenge is to provide high user experienced data rate anytime and anywhere, even for users in remote places with poor infrastructure. Deploying massive infrastructure in these areas is cost inefficient due to the low population density and hostile environment. Hybrid air-terrestrial networks, where high altitude platforms (HAPs) are deployed to provide internet access for ground stations or user equipments (UEs), are promising approach to deal with this problem. HAPs are airships or aircrafts in the stratosphere (at an altitude around 20 km), which can provide ubiquitous wireless access over large coverage areas at low cost. One example of HAP is Google Balloon [1] which uses high altitude balloons to create a hybrid air-terrestrial network with up to LTE data rate. Another example is ''Internet From Sky'' proposed by Facebook. In this project, unmanned aerial vehicles are deployed to provide a novel and efficient method of access [2].

Despite of HAP, the deployment of low altitude platform (LAP) under 1 km for communications has also drawn much attention recently. LAP can be used in the high-capacity hot-spot scenario which is one of the main technical scenario for the fifth generation (5G) mobile communication

systems [3]. In this scenario, LAPs (such as balloons and unmanned aerial vehicles) play the role of temporary base stations (BSs) to provide ultra-high data rates for hot-spot UEs. Another application of LAP is for the wireless recovery networks, or called emergency supplementary networks (ESN) in [4]. The concept of ESN has been accepted by the Homeland Security Bureau in USA, which aims to recover the critical communications for first responders within 12-18 hours by deployable air-bone communication systems [5]. Other application scenarios of LAP includes mobile relaying and information dissemination/data collection in wireless sensor networks (see [6] and the reference therein).

In conventional air-to-ground communication systems, air platform (AP) is usually equipped with directional antennas which can generate several spot beams from AP for data transmission [7]. Due to limited spatial resolution of directional antennas, the coverage radius of each spot beam on the ground is up to several kilometers. As a result, no spatial multiplexing gain can be obtained if the UEs located in the coverage region of one spot beam, which limits its system spectral efficiency (SE). Massive multiple-input multiple-output (MIMO) can provide high spatial resolution by employing large-scale antenna arrays [8]–[18], and thus is a promising technology to address this problem. Additionally, with coherent processing massive MIMO can achieve tremendous power gain, which is beneficial to combat the path loss introduced by both large vertical and horizontal distances. In the terrestrial communication systems, the antenna elements of base station (BS) are usually placed in a line to form uniform linear array (ULA) [11]. As shown in [13], for LTE carrier frequency of 2.5 GHz, it requires about 1.9 m to place a ULA with 32 antenna elements. Using this architecture, it is difficult to equip a large number of antennas at AP due to the physical size of antenna array. In contrast, by placing antenna elements in a two-dimension grid, the planar array architecture enables a large number of antenna elements in a compact area. Moreover, planar array can control radiation pattern in a three-dimension (3D) space (forming a 3D massive MIMO system) and focus more precisely on the intended UEs, which can potentially achieve better power gain [14].

In this paper, we investigate the feasibility of 3D massive MIMO for air-to-ground transmission, where we assume an AP is equipped with a two-dimension (2D) rectangular antenna array and communicates with a number of UEs on the ground. In such system, the biggest challenge is to design practical downlink (air-to-ground) precoding scheme. Different from the BS on the ground, AP is hardware and computational complexity limited due to the consideration of weight, fabricating cost and energy supply. Traditional linear precoding schemes, such as match filter precoding and zero-forcing (ZF) precoding [15], [16], require a large number radio frequency (RF) chains at AP and have high computational complexity. This may make these schemes infeasible when applied on AP. The recent basis expansion based precoding can reduce the number of required RF chains significantly by exploiting the compressibility of channels in beam domain [17]. However, as will be shown in the simulation, this scheme does not perform well when the angular spreads of signals are extremely narrow (which occurs when the altitude of AP increases) and the number of RF chains is very small.

This paper studies low-complexity 3D massive MIMO precoding for air-to-ground transmission considering the above challenges. The contributions are summarized as follows.

- By exploiting the slow time-varying parameters, such as channel correlation and angles of departure (AoDs) of UEs, we first propose a location-assisted two-layer precoding scheme for downlink transmission. Then by proper UE clustering, we approximate the first-layer precoding matrix using a constant-envelope matrix, which can be realized using phase shifting networks with low hardware complexity. With the proposed precoding scheme, we analyze the effect of AP's altitude on the hardware design of AP, and show that the required number of RF chains can be reduced when the altitude of AP increases.
- As the AoD information is usually not perfectly known at AP due to estimation error and relative movement between AP and UEs, we investigate the impact of AoD uncertainty on the location-assisted precoding. Particularly, we propose a new analytic method to estimate the power loss due to AoD errors, which is very useful in link budget and design of AoD estimation scheme.
- At last, we present numerical simulations to evaluate performance of location-assisted precoding under different system parameters, including Rician factor, AP's altitude and AoD uncertainty. Our results show that the proposed scheme outperforms match filter and basis expansion based precoding in the air-to-ground transmission scenarios significantly.

The rest of the paper is organized as follows. Section II reviews the related work. Section III presents the system and channel models. Section IV presents the location-assisted precoding scheme. Section IV analyzes the effect of AoD uncertainty. Section V presents the simulation results. Some conclusions are drawn at last in section VI.

Notations: The symbol $\jmath$ denotes $\sqrt{-1}$. $\delta(\cdot)$ denotes the dirac delta function. $\mathbb{E}(\cdot)$ denotes the expectation. $\mathbf{I}_n$ denotes the $n \times n$ identity matrix. $\otimes$ denotes the Kronecker product. $(\cdot)^T$, $(\cdot)^H$, $\mathrm{tr}(\cdot)$ and $||\cdot||$ denote the transpose, conjugate-transpose, trace and Euclidean norm of matrix, respectively. $\mathrm{eig}_n(\mathbf{A})$ denotes the matrix whose columns consist of the first n dominant eigenvectors of $\mathbf{A}$. The distance between two sets B_1 and B_2 with real value elements is defined as $D(B_1, B_2) = \inf_{b_1 \in B_1, b_2 \in B_2} |b_1 - b_2|$. $\mathrm{asinc}_N(x)$ is aliased sinc function, which is defined as $\mathrm{asinc}_N(x) = \frac{\sin(N\pi x)}{N\sin(\pi x)}$. ${}_2F_1(\cdot,\cdot,\cdot)$ denotes hypergeometric function [19, 9.111].

II. RELATED WORK

The recent research on AP for mobile communication focuses on modeling the air-to-ground channels and interferences.

Al-Hourani *et al.* [20] proposed a statistical propagation model to predict the air-to-ground path loss between AP and ground UEs based on ray tracing simulation. Michailidis *et al.* [21] investigated the 3D modeling of small-scale fading for air-to-ground channels, where both line-of-sight (LoS) and non-LoS (NLoS) channel components were considered. Lian *et al.* [22] employed a birth and death precess to model the 'appear' and 'disappear' of physical scatters, which results in a non-stationary channel model. Hu *et al.* [23] considered the modeling of interference in multi-user air-to-ground 3D MIMO channels. The results showed that the total interference can be approximated using a Beta-mixture distribution if the LoS channel is dominated. Zeng *et al.* [6] presented the basic networking architecture and discussed the major design problems in unmanned aerial vehicle (as LAP) aided hybrid air-terrestrial networks.

The research on 3D MIMO has also drawn much attention in terrestrial communication systems. The 3D channel models for vertical precoding in the on-going 3GPP studies were summarized in [24]. An information-theoretic channel model that supports 3D MIMO system has been proposed in [25]. Based on this model, the authors derived analytical expression for the cumulative density function of mutual information. The horizontal and vertical precoding using statistical channel for 3D massive MIMO system have been studied in [14] and [26] by assuming the channel correlation matrix has Kronecker structure. The class of two-layer precoding for 3D MIMO system has been investigated in [11], [17], [27], and [28]. Ayach *et al.* [27] proposed a sparse precoding scheme for point-to-point millimeter wave 3D MIMO system. In this scheme, the first-layer (constant-envelope) and second-layer precoding matrices are jointly designed to approximate the optimal precoding scheme using the orthogonal matching pursuit algorithm. In the schemes of [11] and [17], to satisfy constant-envelope requirement, the authors designed the first-layer precoder as the submatrix of a discrete Fourier transform (DFT) matrix or Kronecker product of DFT matrices. Note that different from [11] and [17], in our scheme the first-layer precoding is not limited to the set of DFT vectors, which can achieves better performance as will be shown in the simulations. In [28], a novel path division multiplexing transmission scheme was proposed for 3D MIMO system with lens antenna array. It was shown that the scheme has remarkable advantages in hardware and signal processing complexity over traditional rectangular array design. Therefore, the application of lens antenna array in air-to-ground transmission is also an interesting and promising research direction in the future.

III. SYSTEM AND CHANNEL MODELS

Consider a network with an AP and L single-antenna UEs on the ground. The AP is equipped with a 2D rectangular antenna array on the x-y plane with N_x antenna elements along x axis and N_y antenna elements along y axis, as shown in Fig. 1. We consider downlink transmission where AP transmits signals to UEs. The received signals at UEs can be

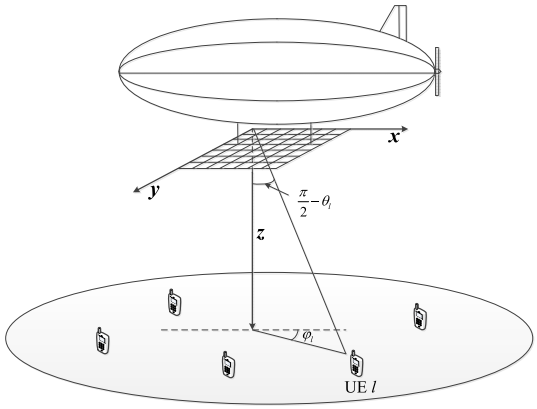


FIGURE 1. Illustration of 3D massive MIMO system for air-to-ground transmission.

expressed as

$$\mathbf{y} = \mathbf{H}^H \mathbf{x} + \mathbf{n}, \tag{1}$$

where $\mathbf{x} \in \mathbb{C}^{N_x N_y \times 1}$ denotes the signal vector transmitted by AP. $\mathbf{n} \in \mathbb{C}^{L \times 1}$ denotes the additive white Gaussian noise (AWGN) vector with distribution $\mathcal{CN}(0, \sigma^2 \mathbf{I}_L)$. $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_L] \in \mathbb{C}^{N_x N_y \times L}$ and $\mathbf{h}_l$ denotes the channel vector from AP to UE l. Based on the air-to-ground 3D MIMO channel model developed in [21], $\mathbf{h}_l$ can be modeled as

$$\mathbf{h}_l = \mathbf{h}_l^{\mathrm{LoS}} + \mathbf{h}_l^{\mathrm{NLoS}}. \tag{2}$$

In (2), $\mathbf{h}_l^{\mathrm{LoS}}$ denotes the deterministic LoS channel component and $\mathbf{h}_l^{\mathrm{NLoS}}$ is the random vector which characterizes the NLoS channel component, which are given by

$$\begin{aligned}
\mathbf{h}_l^{\mathrm{LoS}} &= \sqrt{\frac{\beta_l K_l}{K_l + 1}} \mathbf{b}\left(\varphi_l, \theta_l\right) \otimes \mathbf{a}\left(\varphi_l, \theta_l\right), \\
\mathbf{h}_l^{\mathrm{NLoS}} &= \sqrt{\frac{\beta_l}{K_l + 1}} \int_{\varphi_l - \Delta\varphi_l}^{\varphi_l + \Delta\varphi_l} \int_{\theta_l - \Delta\theta_l}^{\theta_l + \Delta\theta_l} r_l\left(\varphi, \theta\right) \mathbf{b}\left(\varphi, \theta\right) \\
&\quad \otimes \mathbf{a}\left(\varphi, \theta\right) \mathrm{d}\varphi \mathrm{d}\theta,
\end{aligned} \tag{3}$$

with

$$\begin{aligned}
\mathbf{a}\left(\varphi, \theta\right) &= \left[1, \exp\left(\jmath \frac{2\pi d_y}{\lambda} \cos\theta \cos\varphi\right), \cdots, \right. \\
&\quad \left. \exp\left(\jmath \frac{2\pi d_y \left(N_y - 1\right)}{\lambda} \cos\theta \cos\varphi\right)\right]^T, \\
\mathbf{b}\left(\varphi, \theta\right) &= \left[1, \exp\left(\jmath \frac{2\pi d_x}{\lambda} \cos\theta \sin\varphi\right), \cdots, \right. \\
&\quad \left. \exp\left(\jmath \frac{2\pi d_x \left(N_x - 1\right)}{\lambda} \cos\theta \sin\varphi\right)\right]^T,
\end{aligned} \tag{4}$$

where β_l and K_l denote the large-scale fading and Rician factor, respectively. $\varphi_l \in [-\pi/2, \pi/2]$ and $\theta_l \in (0, \pi/2]$ denote the AoDs of UE l in horizontal and vertical directions, and $\Delta\varphi_l$ and $\Delta\theta_l$ denote the corresponding angular spreads.

d_x and d_y denote the antenna spacings along the x axis and y axis respectively, and λ denotes the carrier wavelength. $r_l(\theta, \varphi)$ denotes the complex-valued response gain associated to the direction (θ, φ). We assume that $r_l(\theta, \varphi)$ has zero-mean, i.e., $\mathbb{E}[r_l(\theta, \varphi)] = 0$, and the response gains for different incidence angles are uncorrelated, i.e.,

$$\mathbb{E}\left[r_l^*(\varphi,\theta)\, r_l\left(\varphi',\theta'\right)\right] = S_l(\varphi,\theta)\,\delta\left(\varphi-\varphi'\right)\delta\left(\theta-\theta'\right), \tag{5}$$

where $S_l(\varphi, \theta)$ represents the channel power angle spectrum (PAS) which characterizes the channel power distribution in angular domain. According to the model in (2)-(5), the channel correlation matrix can be expressed as

$$\begin{aligned}
\mathbf{C}_l &= \mathbb{E}\left[\mathbf{h}_l\mathbf{h}_l^H\right] \\
&= \mathbb{E}\left[\mathbf{h}_l^{\mathrm{LoS}}\left(\mathbf{h}_l^{\mathrm{LoS}}\right)^H\right] + \mathbb{E}\left[\mathbf{h}_l^{\mathrm{NLoS}}\left(\mathbf{h}_l^{\mathrm{NLoS}}\right)^H\right] \\
&= \frac{\beta_l K_l}{K_l+1}\left(\mathbf{b}(\varphi_l,\theta_l)\,\mathbf{b}^H(\varphi_l,\theta_l)\right)\otimes\left(\mathbf{a}(\varphi_l,\theta_l)\,\mathbf{a}^H(\varphi_l,\theta_l)\right) \\
&\quad + \frac{\beta_l}{K_l+1}\int_{\varphi_l-\Delta\varphi_l}^{\varphi_l+\Delta\varphi_l}\int_{\theta_l-\Delta\theta_l}^{\theta_l+\Delta\theta_l} S_l(\varphi,\theta)\left(\mathbf{b}(\varphi,\theta)\,\mathbf{b}^H(\varphi,\theta)\right) \\
&\quad \otimes\left(\mathbf{a}(\varphi,\theta)\,\mathbf{a}^H(\varphi,\theta)\right)\mathrm{d}\varphi\mathrm{d}\theta \\
&\triangleq \mathbf{C}_l^{\mathrm{LoS}} + \mathbf{C}_l^{\mathrm{NLoS}}.
\end{aligned} \tag{6}$$

In the terrestrial MIMO system, the antenna array of BS is usually placed on y-z or x-z plane. Moreover, the AoD in the vertical direction is small due to limited altitude of BS. In this case, the correlation matrix of channel can be written approximately as the Kronecker product of correlation matrices in horizontal and vertical directions [11], [26]. Due to this reason, the horizontal and vertical precoding scheme at BS can be designed separately to reduce the complexity [11], [26]. However, in order to cover the ground UEs, the antenna array of AP is placed on x-y plane. Moreover, as the altitude of AP increases, the AoD in the vertical direction becomes non-negligible. Thus, the assumption of Kronecker structure for correlation matrix of channel does not hold, as observed in (6). This increases the difficulty in precoding design, because a joint but still low-complexity precoding scheme is required.

For further analysis, we find it convenient to define $\rho_x = \frac{d_x}{\lambda}\cos\theta\sin\varphi$ and $\rho_y = \frac{d_y}{\lambda}\cos\theta\cos\varphi$ as the virtual AoDs along x-axis and y-axis, respectively. Then we can give another definition of correlation matrix for NLoS channel

$$\begin{aligned}
\mathbf{C}_l^{\mathrm{NLoS}} &= \frac{\beta_l}{K_l+1}\int_{\rho_{l,x}^{\min}}^{\rho_{l,x}^{\max}}\int_{\rho_{l,y}^{\min}}^{\rho_{l,y}^{\max}} S_l^{\mathrm{v}}(\rho_x,\rho_y)\left(\mathbf{b}(\rho_x)\,\mathbf{b}^H(\rho_x)\right), \\
&\qquad \otimes\left(\mathbf{a}(\rho_y)\,\mathbf{a}^H(\rho_y)\right)\mathrm{d}\rho_x\mathrm{d}\rho_y \\
\mathbf{a}(\rho_y) &= \left[1, \exp\left(\jmath 2\pi\rho_y\right), \cdots, \exp\left(\jmath 2\pi\left(N_y-1\right)\rho_y\right)\right]^T, \\
\mathbf{b}(\rho_x) &= \left[1, \exp\left(\jmath 2\pi\rho_x\right), \cdots, \exp\left(\jmath 2\pi\left(N_x-1\right)\rho_x\right)\right]^T.
\end{aligned} \tag{7}$$

The integral boundaries are given by

$$\begin{aligned}
\rho_{l,x}^{\max} &= \max_{\theta\in[\theta_l-\Delta\theta_l,\theta_l+\Delta\theta_l],\varphi\in[\varphi_l-\Delta\varphi_l,\varphi_l+\Delta\varphi_l]} \frac{d_x}{\lambda}\cos(\theta)\sin(\varphi), \\
\rho_{l,x}^{\min} &= \min_{\theta\in[\theta_l-\Delta\theta_l,\theta_l+\Delta\theta_l],\varphi\in[\varphi_l-\Delta\varphi_l,\varphi_l+\Delta\varphi_l]} \frac{d_x}{\lambda}\cos(\theta)\sin(\varphi), \\
\rho_{l,y}^{\max} &= \max_{\theta\in[\theta_l-\Delta\theta_l,\theta_l+\Delta\theta_l],\varphi\in[\varphi_l-\Delta\varphi_l,\varphi_l+\Delta\varphi_l]} \frac{d_y}{\lambda}\cos(\theta)\cos(\varphi), \\
\rho_{l,y}^{\min} &= \min_{\theta\in[\theta_l-\Delta\theta_l,\theta_l+\Delta\theta_l],\varphi\in[\varphi_l-\Delta\varphi_l,\varphi_l+\Delta\varphi_l]} \frac{d_y}{\lambda}\cos(\theta)\cos(\varphi).
\end{aligned} \tag{8}$$

Moreover, $S_l^{\mathrm{v}}(\rho_x, \rho_y)$ denotes the PAS with respect to ρ_x and ρ_y, which can be derived using the formula of integration by substitution as

$$\begin{aligned}
&S_l^{\mathrm{v}}(\rho_x,\rho_y) \\
&= S_l\left(\arctan\left(\frac{d_y\rho_x}{d_x\rho_y}\right), \arccos\left(\lambda\sqrt{\left(\frac{\rho_x}{d_x}\right)^2+\left(\frac{\rho_y}{d_y}\right)^2}\right)\right) \\
&\quad\times\frac{\lambda^2}{d_x d_y}\sqrt{\frac{1}{1-\left(\frac{\lambda\rho_x}{d_x}\right)^2-\left(\frac{\lambda\rho_y}{d_y}\right)^2}\frac{1}{\left(\frac{\lambda\rho_x}{d_x}\right)^2+\left(\frac{\lambda\rho_y}{d_y}\right)^2}}.
\end{aligned} \tag{9}$$

In (3), the channel modeling is in fact based on the well-known ray-tracing approach. The NLoS channel component can be viewed as the sum of a large number of physical paths with different horizontal and vertical AoDs. Therefore, using the law of large numbers, we assume the NLoS channel component $\mathbf{h}_l^{\mathrm{NLoS}}$ has correlated Gaussian distribution with zero-mean and covariance matrix $\mathbf{C}_l^{\mathrm{NLoS}}$. Similar assumption has also been used in [29].

IV. DOWNLINK PRECODING DESIGN

In this section, we first propose a location-assisted two-layer precoding scheme for downlink transmission. Then we show that, by clustering the UEs properly, the first-layer precoding matrix can be approximated using a constant-envelope matrix, which can be realized with phase shifting networks in the analog domain [31]. This results in a significant reduction on hardware complexity of AP.

A. LOCATION-ASSISTED PRECODING

We assume that the UEs are divided into C clusters. The number of UEs in cluster $c \in \{1, 2, \cdots, C\}$ is L_c which satisfies $\sum_{c=1}^{C} L_c = L$. We define $c_l = l + \sum_{c'=1}^{c-1} L_{c'}$ as the index of the lth UE in cluster c. The AP serves all clusters using the same time-frequency resource. Hence, we design the transmit signal vector of AP as $\mathbf{x} = \sum_{c=1}^{C}\mathbf{x}_c$, where $\mathbf{x}_c \in \mathbb{C}^{N_xN_y\times 1}$ denotes the signal transmitted to cluster c.

The proposed scheme consists of two-layer precoding. Defining $\mathbf{P}_c \in \mathbb{C}^{N_xN_y\times b_c}$ and $\mathbf{W}_c \in \mathbb{C}^{b_c\times L_c}$ as the first- and second-layer precoding matrices, we can express $\mathbf{x}_c$ as $\mathbf{x}_c = \mathbf{P}_c\mathbf{W}_c\mathbf{s}_c$ where $\mathbf{s}_c$ is a $L_c\times 1$ vector contains the symbols intended at cluster c. Note that b_c is a design parameter which determines the dimension of small MIMO system and satisfies $L_c \le b_c \ll N_x\times N_y$. At last, we define $\mathbf{H}_c \in \mathbb{C}^{N_xN_y\times L_c}$ as

the channel matrix between the AP and UE cluster c. Using the above definitions, the signal model (1) can be rewritten as

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_C \end{bmatrix} = \begin{bmatrix} \mathbf{H}_1^H \sum_{c=1}^{C} \mathbf{P}_c \mathbf{W}_c \mathbf{s}_c \\ \mathbf{H}_2^H \sum_{c=1}^{C} \mathbf{P}_c \mathbf{W}_c \mathbf{s}_c \\ \vdots \\ \mathbf{H}_C^H \sum_{c=1}^{C} \mathbf{P}_c \mathbf{W}_c \mathbf{s}_c \end{bmatrix} + \begin{bmatrix} \mathbf{n}_1 \\ \mathbf{n}_2 \\ \vdots \\ \mathbf{n}_C \end{bmatrix}, \quad (10)$$

where $\mathbf{n}_c \in \mathbb{C}^{L_c \times 1}$ and $\mathbf{y}_c \in \mathbb{C}^{L_c \times 1}$ are the AWGN and received signal vector associated with the cluster c. For further analysis, let us rewrite $\mathbf{y}_c \in \mathbb{C}^{L_c \times 1}$ as follows

$$\mathbf{y}_c = \mathbf{H}_c^H \mathbf{P}_c \mathbf{W}_c \mathbf{s}_c + \mathbf{H}_c^H \sum_{c'=1, c' \neq c}^{C} \mathbf{P}_{c'} \mathbf{W}_{c'} \mathbf{s}_{c'} + \mathbf{n}_c. \quad (11)$$

The first term of right-hand side of (11) is the intended signal of cluster c and the second term is the inter-cluster interference (ICI).

1) FIRST-LAYER PRECODING

The objective of first-layer precoding is to decompose the original massive MIMO system into several small MIMO systems, with each operating on the orthogonal subspace. To this end, we design $\mathbf{P}_c$ to eliminate the ICI using the correlation knowledge of channels (which can be obtained by exploiting the location information of UEs, including AoDs and angular spreads according to (6)). The benefits are two-fold: 1) For the decomposed systems, the total required length of training sequence to obtain the channel state information (CSI) can be substantially reduced. 2) As will be shown in the next subsection, with proper UE clustering we can approximate the first-layer precoder using a constant-envelope matrix. This can substantially reduce the hardware complexity at AP. In addition to ICI cancellation, we also wish that the average output signal power after the first-layer precoding is as large as possible. Thus, the optimal $\mathbf{P}_c$ can be determined by solving the following problem

$$\max_{\mathbf{P}_c} \ \mathbb{E}\left[\left\| \mathbf{H}_c^H \mathbf{P}_c \mathbf{W}_c \mathbf{s}_c \right\|^2 \right] \quad (12a)$$

$$\text{s.t.} \ \left\| \mathbf{H}_{c'}^H \mathbf{P}_c \mathbf{W}_c \mathbf{s}_c \right\|^2 = 0, \quad \forall c' \neq c, \quad (12b)$$

$$\mathbf{P}_c^H \mathbf{P}_c = \mathbf{I}_{b_c}. \quad (12c)$$

From (12a), the design of $\mathbf{P}_c$ is coupled with the second-layer precoding matrix, which makes closed-form solution inaccessible. Thus, it is of great interest to find new target function to decouple the problem. To this end, by using Cauchy-Schwarz inequality, an upper bound on average output power can be obtained as

$$\begin{aligned} \mathbb{E}\left[\left\| \mathbf{H}_c^H \mathbf{P}_c \mathbf{W}_c \mathbf{s}_c \right\|^2 \right] &\leq \mathbb{E}\left[\mathrm{tr}\left(\mathbf{P}_c^H \mathbf{H}_c \mathbf{H}_c^H \mathbf{P}_c \right) \right] \mathbb{E}\left[\| \mathbf{W}_c \mathbf{s}_c \|^2 \right] \\ &= \sum_{l=1}^{L_c} \mathrm{tr}\left(\mathbf{P}_c^H \mathbf{C}_{c_l} \mathbf{P}_c \right) \mathbb{E}\left[\| \mathbf{W}_c \mathbf{s}_c \|^2 \right]. \end{aligned} \quad (13)$$

In the following, we define $\sum_{l=1}^{L_c} \mathrm{tr}\left(\mathbf{P}_c^H \mathbf{C}_{c_l} \mathbf{P}_c \right)$ as the power gain of the first-layer precoding. Then the original problem becomes finding $\mathbf{P}_c$ that maximizes the power gain subject to the ICI cancellation constraint.

On the other hand, to simplify the constraint (12b), we present the following lemma.

Lemma 1: A sufficient condition for constraint (12b) can be expressed as

$$\sum_{c'=1, c' \neq c}^{C} \sum_{l=1}^{L_c} \mathbf{P}_c^H \mathbf{C}_{c'_l} \mathbf{P}_c = 0. \quad (14)$$

Proof: Note that a sufficient condition for (12b) is

$$\left| \mathbf{h}_{c'_l}^H \mathbf{P}_c \right|^2 = 0, \quad \forall c' \neq c, \ \forall l \in \{1, 2, \cdots, L_{c'}\}. \quad (15)$$

According to the channel model in section III, (15) can be expanded as

$$\begin{aligned} \left| \mathbf{h}_{c'_l}^H \mathbf{P}_c \right|^2 &= \mathrm{tr}\left(\mathbf{P}_c^H \mathbf{C}_{c'_l}^{\mathrm{LoS}} \mathbf{P}_c \right) \\ &\quad + \mathrm{tr}\left(\mathbf{P}_c^H \left(\mathbf{C}_{c'_l}^{\mathrm{NLoS}} \right)^{1/2} \mathbf{q}\mathbf{q}^H \left(\mathbf{C}_{c'_l}^{\mathrm{NLoS}} \right)^{1/2} \mathbf{P}_c \right) \\ &\quad + 2\mathrm{Re}\left\{ \mathrm{tr}\left(\mathbf{P}_c^H \mathbf{h}_{c'_l}^{\mathrm{LoS}} \mathbf{q}^H \left(\mathbf{C}_{c'_l}^{\mathrm{NLoS}} \right)^{1/2} \mathbf{P}_c \right) \right\} = 0, \end{aligned} \quad (16)$$

where $\mathbf{q} \sim \mathcal{CN}(0, \mathbf{I}_{N_x N_y})$. Note that second and third terms of (16) are zero as long as $\mathbf{P}_c^H \left(\mathbf{C}_{c'_l}^{\mathrm{NLoS}} \right)^{1/2} = 0$. According to the definition in (6), $\mathbf{C}_{c'_l}^{\mathrm{NLoS}}$ is a positive semi-definite matrix, and $\mathbf{P}_c^H \left(\mathbf{C}_{c'_l}^{\mathrm{NLoS}} \right)^{1/2} = 0$ is equivalent to $\mathbf{P}_c^H \left(\mathbf{C}_{c'_l}^{\mathrm{NLoS}} \right)^{1/2} \mathbf{P}_c = 0$. Therefore, to make (16) hold, it is sufficient to let

$$\begin{aligned} \mathbf{P}_c^H \mathbf{C}_{c'_l}^{\mathrm{LoS}} \mathbf{P}_c &= 0, \\ \mathbf{P}_c^H \left(\mathbf{C}_{c'_l}^{\mathrm{NLoS}} \right)^{1/2} \mathbf{P}_c &= 0. \end{aligned} \quad (17)$$

Note that $\left(\mathbf{C}_{c'_l}^{\mathrm{NLoS}} \right)^{1/2}$ and $\mathbf{C}_{c'_l}^{\mathrm{NLoS}}$ have the same null space. Thus, the second equation in (17) can be replaced by $\mathbf{P}_c^H \mathbf{C}_{c'_l}^{\mathrm{NLoS}} \mathbf{P}_c = 0$. At last, according to the positive semi-definition of $\mathbf{C}_{c'_l}^{\mathrm{LoS}}$ and $\mathbf{C}_{c'_l}^{\mathrm{NLoS}}$, the equality (17), $\forall c' \neq c, \forall l \in \{1, 2, \cdots, L_{c'}\}$, is equivalent to (14). ■

Using (13) and Lemma 1, we can rewrite problem (12) as

$$\max_{\mathbf{P}_c} \ \sum_{l=1}^{L_c} \mathrm{tr}\left(\mathbf{P}_c^H \mathbf{C}_{c_l} \mathbf{P}_c \right) \quad (18a)$$

$$\text{s.t.} \ \sum_{c'=1, c' \neq c}^{C} \sum_{l=1}^{L_c} \mathbf{P}_c^H \mathbf{C}_{c'_l} \mathbf{P}_c = 0, \quad (18b)$$

$$\mathbf{P}_c^H \mathbf{P}_c = \mathbf{I}_{b_c}. \quad (18c)$$

Let $\mathbf{V}_{(-c)}$ be the matrix contains all eigenvectors of $\sum_{c'=1, c' \neq c}^{C} \sum_{l=1}^{L_c} \mathbf{C}_{c'_l}$ with zero eigenvalue. Then the solution of problem (18) that satisfies constraints (18b) and (18c) can be expressed as $\mathbf{P}_c = \mathbf{V}_{(-c)} \mathbf{U}_{(-c)}$, where $\mathbf{U}_{(-c)}$ is a weighting

matrix with $\mathbf{U}_{(-c)}^H\mathbf{U}_{(-c)} = \mathbf{I}_{b_c}$. Therefore, the problem (18) is equivalent to finding the optimal $\mathbf{U}_{(-c)}$, which can be expressed as the following problem

$$\max_{\mathbf{U}_{(-c)}^H\mathbf{U}_{(-c)}=\mathbf{I}_{b_c}} \sum_{l=1}^{L_c} \mathrm{tr}\left(\mathbf{U}_{(-c)}^H\mathbf{V}_{(-c)}^H\mathbf{C}_{c_l}\mathbf{V}_{(-c)}\mathbf{U}_{(-c)}\right). \quad (19)$$

It is easy to shown that the optimal solution is

$$\mathbf{U}_{(-c)} = \mathrm{eig}_{b_c}\left(\sum_{l=1}^{L_c}\mathbf{V}_{(-c)}^H\mathbf{C}_{c_l}\mathbf{V}_{(-c)}\right). \quad (20)$$

Note that in the first-layer precoding described in the above, we have not added any constraints on UE clustering. In fact, since the angular spreads in both horizontal and vertical direction are narrow for far-field transmission, the effective rank of channel correlation matrix are much smaller than the dimension of AP antenna array (i.e., $N_x \times N_y$) and $\sum_{c'=1,c'\neq c}^{C}\sum_{l=1}^{L_c}\mathbf{C}_{c'l}$ has large enough null-space [17]. As a result, for arbitrary UE clustering we can always find the solution of $\mathbf{P}_c$ since $b_c \ll N_x \times N_y$.

2) SECOND-LAYER PRECODING

The second-layer precoding is to eliminate the multi-UE interference within the cluster. After first-layer precoding, the received signal at cluster c given by (11) becomes

$$\mathbf{y}_c = \mathbf{H}_c^H\mathbf{P}_c\mathbf{W}_c\mathbf{s}_c + \mathbf{n}_c. \quad (21)$$

As a result, $\mathbf{W}_c$ can be designed as the well-known ZF precoder, that is

$$\mathbf{W}_c = \mathbf{P}_c^H\mathbf{H}_c\left(\mathbf{H}_c^H\mathbf{P}_c\mathbf{P}_c^H\mathbf{H}_c\right)^{-1}\Gamma_c^{\frac{1}{2}}, \quad (22)$$

where Γ_c is a diagonal normalization matrix with $[\Gamma_c]_{l,l} = p_{c_l}/[(\mathbf{P}_c^H\mathbf{H}_c\mathbf{H}_c^H\mathbf{P}_c)^{-1}]_{l,l}$ and p_{c_l} denotes the power allocated to the signal of UE c_l. Let $\tilde{\mathbf{H}}_c = \mathbf{P}_c^H\mathbf{H}_c$ denote the $b_c \times L_c$ effective channel matrix between AP and cluster c. To implement the second-layer precoding, $\tilde{\mathbf{H}}_c$ should be estimated before downlink transmission. In the proposed scheme, since the signals transmitted to different UE clusters lie in orthogonal subspaces after the first-layer precoding, the same pilot sequence can be reused by all clusters without any pilot contamination. Therefore, the minimum lengthes of training sequence (which is equal to the minimum numbers of required orthogonal pilot sequences) to obtain the effective channel estimates of all clusters are $\max_{c\in\{1,2,\cdots,C\}} L_c$ (symbol time) in time-division duplex (TDD) setting and $\max_{c\in\{1,2,\cdots,C\}} b_c$ (symbol time) in frequency-division duplex (FDD) setting. This results in a significant saving in training resource.[1]

In FDD system, the channel reciprocity does not hold and downlink training is required. After estimating the CSI from the downlink pilot signals, UEs should quantize estimated CSI and then transmit the quantization back to AP through feedback channels. This can affect the system from two aspects. First, the feedback error due to quantization error, noise and feedback delay decreases the accuracy of CSI. Moreover, CSI feedback increases the load of feedback channel, and hence can degrade the overall system SE. Note that the results in [30] have shown that the CSI error (in term of mean-square error) due to feedback can be made much smaller than that caused by estimation error in downlink training phase. Moreover, since the effective downlink channel dimension (after the first-layer precoding) is greatly reduced, we assume that the additional load caused by CSI feedback is small compared to other feedback information. Therefore, for simplicity, we consider the optimistic situation of error-free CSI feedback and neglect the SE penalty due to feedback. A similar approach is also adopted in analysis of [11].

[1]It is well-known that, in the precoding schemes without UEs' location information [17], the minimum required length of training sequence to obtain the CSI of all UEs is $\sum_{c=1}^{C} L_c$ (symbol time) in the TDD setting and $N_x \times N_y$ (symbol time) in the FDD setting.

B. UE CLUSTERING AND REDUCED COMPLEXITY PRECODING

To implement the location-assisted precoding in the last subsection, AP should be equipped with $N_x \times N_y$ RF chains. This becomes quite challenging when N_x and N_y grow very large, since AP is hardware complexity limited due to the consideration of weight, fabricating cost and energy supply. In this subsection, with proper UE clustering, we show that the first-layer precoding matrix can be approximated by a constant-envelope matrix which can be realized with phase shifting networks in the analog domain [31]. In this way, the total number of required RF chains reduces to $\sum_{c=1}^{C} b_c$.

1) UE CLUSTERING

The UE clustering is based on the following lemma.

Lemma 2: Consider the unitary and constant-envelope vector $\mathbf{g}(\omega_x,\omega_y) = \tilde{\mathbf{g}}_{N_x}(\omega_x)\otimes\tilde{\mathbf{g}}_{N_y}(\omega_y)$, where $\tilde{\mathbf{g}}_{N_x}(\omega_x)$ and $\tilde{\mathbf{g}}_{N_y}(\omega_y)$ are defined as $\tilde{\mathbf{g}}_{N_x}(\omega_x) = \frac{1}{\sqrt{N_x}}[1, \exp(\jmath 2\pi\omega_x), \cdots, \exp(\jmath 2\pi(N_x-1)\omega_x)]^T$ and $\tilde{\mathbf{g}}_{N_y}(\omega_y) = \frac{1}{\sqrt{N_y}}[1, \exp(\jmath 2\pi\omega_y), \cdots, \exp(\jmath 2\pi(N_y-1)\omega_y)]^T$. As N_x and N_y tend to infinity, the quantity $\mathbf{g}^H(\omega_x,\omega_y)\mathbf{C}_{c_l}\mathbf{g}(\omega_x,\omega_y)$ converges to zero if $\omega_x \notin [\rho_{c_l,x}^{\min}, \rho_{c_l,x}^{\max}]$ and $\omega_y \notin [\rho_{c_l,y}^{\min}, \rho_{c_l,y}^{\max}]$. If the NLoS channel component is dominated, i.e., $K_{c_l} \approx 0$, $\mathbf{g}^H(\omega_x,\omega_y)\mathbf{C}_{c_l}\mathbf{g}(\omega_x,\omega_y)$ converges to zero (as $N_x, N_y \to \infty$) if $\omega_x \notin [\rho_{c_l,x}^{\min}, \rho_{c_l,x}^{\max}]$ or $\omega_y \notin [\rho_{c_l,y}^{\min}, \rho_{c_l,y}^{\max}]$.

Proof: The proof has been presented in [33]. ■

According to Lemma 2, if UEs are divided so that the feasible regions of virtual AoDs for different clusters, are non-overlapping, that is, $\bigcap_{c=1}^{C}\left(\bigcup_{l=1}^{L_c}[\rho_{c_l,x}^{\min}, \rho_{c_l,x}^{\max}]\right) = \emptyset$ and $\bigcap_{c=1}^{C}\left(\bigcup_{l=1}^{L_c}[\rho_{c_l,y}^{\min}, \rho_{c_l,y}^{\max}]\right) = \emptyset$, and the first-layer precoding matrix has the following structure

$$\mathbf{P}_c = [\mathbf{g}(\omega_{x,1},\omega_{y,1}), \mathbf{g}(\omega_{x,2},\omega_{y,2}), \cdots, \mathbf{g}(\omega_{x,b_c},\omega_{y,b_c})], \quad (23a)$$

$$\text{with} \begin{cases} \omega_{x,1}, \omega_{x,2}, \cdots, \omega_{x,b_c} \in \bigcup_{l=1}^{L_c} \left[\rho_{c_l,x}^{\min}, \rho_{c_l,x}^{\max}\right] \\ \omega_{y,1}, \omega_{y,2}, \cdots, \omega_{y,b_c} \in \bigcup_{l=1}^{L_c} \left[\rho_{c_l,y}^{\min}, \rho_{c_l,y}^{\max}\right], \end{cases} \tag{23b}$$

the ICI constraint in (18b) is satisfied automatically in the large $\{N_x, N_y\}$ regime. Thus, we have the following UE clustering criterion.

UE Clustering Criterion: The UEs are divided so that the feasible regions of virtual AoDs for different clusters satisfy $D\left(\bigcup_{l=1}^{L_c}\left[\rho_{c_l,x}^{\min}, \rho_{c_l,x}^{\max}\right], \bigcup_{l=1}^{L_{c'}}\left[\rho_{c'_l,x}^{\min}, \rho_{c'_l,x}^{\max}\right]\right) \geq G$ and $D\left(\bigcup_{l=1}^{L_c}\left[\rho_{c_l,y}^{\min}, \rho_{c_l,y}^{\max}\right], \bigcup_{l=1}^{L_{c'}}\left[\rho_{c'_l,y}^{\min}, \rho_{c'_l,y}^{\max}\right]\right) \geq G, \forall c, c' \in \{1, 2, \cdots, C\}$ and $c \neq c'$, where G denotes the guard interval.

Lemma 3: If the above criterion is satisfied, the ICI powers received at UE c_l from the LoS and NLoS channels scale with $\mathcal{O}\left(\frac{1}{N_x N_y G^2}\right)$ and $\mathcal{O}\left(\prod_{i\in\{x,y\}} \frac{\rho_{c_l,i}^{\max} - \rho_{c_l,i}^{\max}}{N_i \pi^2 G\left(G + \rho_{c_l,i}^{\max} - \rho_{c_l,i}^{\max}\right)}\right)$, respectively.

Proof: This can be deduced from [33, eq. (23) and eq. (25)]. ■

Lemma 3 indicates that we can arrange UE cluster denser when the number of AP antennas increases or the angular spreads decrease. Otherwise, we must increase the guard interval to control ICI. Note that if NLoS channel is dominated, the constraint on feasible regions of virtual AoDs in UE clustering criterion can be relaxed as $D\left(\bigcup_{l=1}^{L_c}\left[\rho_{c_l,x}^{\min}, \rho_{c_l,x}^{\max}\right], \bigcup_{l=1}^{L_{c'}}\left[\rho_{c'_l,x}^{\min}, \rho_{c'_l,x}^{\max}\right]\right) \geq G$ or $D\left(\bigcup_{l=1}^{L_c}\left[\rho_{c_l,y}^{\min}, \rho_{c_l,y}^{\max}\right], \bigcup_{l=1}^{L_{c'}}\left[\rho_{c'_l,y}^{\min}, \rho_{c'_l,y}^{\max}\right]\right) \geq G$ according to Lemma 2.

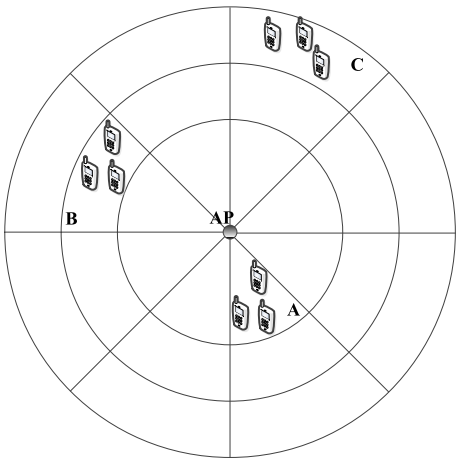


FIGURE 2. Illustration of network-centric UE clustering.

In practical implementation, the UE clustering can be performed based on the network-centric approach or UE-centric approach. In the network-centric approach, the coverage area of AP is divided into a number of small regions according to the horizontal and vertical AoDs with respect to AP, as shown in Fig. 2. The UEs lie in the same small region are collected in the same cluster. To meet the UE clustering criterion, in each time-frequency resource unit, AP picks several UE clusters in the small regions with different horizontal and vertical AoDs (for example, the three UE clusters in regions A, B and C in Fig. 2) to serve. In the simulations of section VI, this approach is adopted for UE clustering. In the UE centric approach, the UE clustering is implemented based on similarity between UEs' horizontal and vertical AoDs or virtual AoDs. In particular, the UEs with similar AoDs are first gathered into the same cluster. This task can be completed using the method based on K-mean principle developed in [32]. To meet the UE clustering criterion in section IV-B, in each time-frequency resource unit, AP picks several UE clusters with disjoint AoDs region to serve.

Another problem is how to choose the number of clusters C. In general, C should be designed to optimize some performance metrics, such as the SE or SE per UE. Due to the space limitation, this problem will not be investigated in the current work. In the following, we consider fixed C for simplicity.

2) REDUCED COMPLEXITY PRECODING

Using the UE clustering criterion and assuming $\mathbf{P}_c$ has the structure in (23), the design of first-layer precoding becomes the problem of finding optimal $\{\omega_{x,i}\}_{i=1}^{b_c}$ and $\{\omega_{y,i}\}_{i=1}^{b_c}$, that is

$$\begin{aligned} &\{\omega_{x,1}, \cdots \omega_{x,b_c}; \omega_{y,1}, \cdots, \omega_{y,b_c}\} \\ &= \arg \max_{\omega_{x,1},\cdots\omega_{x,b_c};\omega_{y,1},\cdots,\omega_{y,b_c}} \sum_{l=1}^{L_c} \mathrm{tr}\left(\mathbf{P}_c^H \mathbf{C}_{c_l} \mathbf{P}_c\right), \\ &\quad \text{s.t.} \begin{cases} (23b), \\ \mathbf{P}_c^H \mathbf{P}_c = \mathbf{I}_{b_c}. \end{cases} \end{aligned} \tag{24}$$

The problem is difficult to solve since the target function has a complex structure with respect to $\{\omega_{x,i}\}_{i=1}^{b_c}$ and $\{\omega_{y,i}\}_{i=1}^{b_c}$. Therefore, in the following we focus on the suboptimal solution. Before the general solution, we first consider the special case with $b_c = L_c$. In this setup, we can rewrite the target function of (24) as

$$\begin{aligned} \sum_{l=1}^{L_c} \mathrm{tr}\left(\mathbf{P}_c^H \mathbf{C}_{c_l} \mathbf{P}_c\right) &= \sum_{l=1}^{L_c} \mathbf{p}_{c,l}^H \mathbf{C}_{c_l} \mathbf{p}_{c,l} \\ &\quad + \sum_{l=1}^{L_c} \mathbf{p}_{c,l}^H \left(\sum_{l'=1, l'\neq l}^{L_c} \mathbf{C}_{c_{l'}}\right) \mathbf{p}_{c,l}. \end{aligned} \tag{25}$$

If we assume that the lth column of $\mathbf{P}_c$, i.e., $\mathbf{p}_{c,l}$, is allocated to transmit the signal of UE c_l, then the lth term in the first summation of (25) can be interpreted as its power gain at c_l, and the lth term in the second summation can be interpreted as the power gain leaked to other UEs $\{c_{l'}\}_{l'\neq l}$. Here we design $\mathbf{p}_{c,l}$ to maximize the power gain at c_l without considering the leaked power, i.e.,

$$\{\omega_{x,l}, \omega_{y,l}\} = \arg \max_{\omega_{x,l},\omega_{y,l}} \mathbf{p}_{c,l}^H \mathbf{C}_{c_l} \mathbf{p}_{c,l}, \quad \text{s.t., } (23b). \tag{26}$$

By substituting (6) and (7) into (26), the target function of (26) can be further expressed as (27), as shown at the top of the next page, where the second step is based

$$\begin{aligned}
\mathbf{p}_{c,l}^{H}\mathbf{C}_{c_l}\mathbf{p}_{c,l} &= \int_{\rho_{l,x}^{\min}}^{\rho_{l,x}^{\max}}\int_{\rho_{l,y}^{\min}}^{\rho_{l,y}^{\max}} S_{c_l}^{\mathrm{v,E}}\left(\rho_x,\rho_y\right)\left(\tilde{\mathbf{g}}_{N_x}\left(\omega_{x,l}\right)\otimes\tilde{\mathbf{g}}_{N_y}\left(\omega_{y,l}\right)\right)^{H}\left(\mathbf{b}\left(\rho_x\right)\mathbf{b}^{H}\left(\rho_x\right)\right)\\
&\quad\otimes\left(\mathbf{a}\left(\rho_y\right)\mathbf{a}^{H}\left(\rho_y\right)\right)\left(\tilde{\mathbf{g}}_{N_x}\left(\omega_{x,l}\right)\otimes\tilde{\mathbf{g}}_{N_y}\left(\omega_{y,l}\right)\right)\mathrm{d}\rho_x\mathrm{d}\rho_y\\
&= \int_{\rho_{l,x}^{\min}}^{\rho_{l,x}^{\max}}\int_{\rho_{l,y}^{\min}}^{\rho_{l,y}^{\max}} S_{c_l}^{\mathrm{v,E}}\left(\rho_x,\rho_y\right)\left|\tilde{\mathbf{g}}_{N_x}^{H}\left(\omega_{x,l}\right)\mathbf{b}\left(\rho_x\right)\right|^2\left|\tilde{\mathbf{g}}_{N_y}^{H}\left(\omega_{y,l}\right)\mathbf{a}\left(\rho_y\right)\right|^2\mathrm{d}\rho_x\mathrm{d}\rho_y\\
&= N_xN_y\int_{\rho_{l,x}^{\min}}^{\rho_{l,x}^{\max}}\int_{\rho_{l,y}^{\min}}^{\rho_{l,y}^{\max}} S_{c_l}^{\mathrm{v,E}}\left(\rho_x,\rho_y\right)\left|\sum_{i=0}^{N_x-1}\exp\left(\jmath 2\pi i\left(\rho_{c_l,x}-\omega_x\right)\right)\right|^2\left|\sum_{i=0}^{N_y-1}\exp\left(\jmath 2\pi i\left(\rho_{c_l,y}-\omega_y\right)\right)\right|^2\mathrm{d}\rho_x\mathrm{d}\rho_y\\
&= N_xN_y\int_{\rho_{c_l,x}^{\min}}^{\rho_{c_l,x}^{\max}}\int_{\rho_{c_l,y}^{\min}}^{\rho_{c_l,y}^{\max}} S_{c_l}^{\mathrm{v,E}}\left(\rho_x,\rho_y\right)\mathrm{asinc}_{N_x}^{2}\left(\omega_{x,l}-\rho_x\right)\mathrm{asinc}_{N_y}^{2}\left(\omega_{y,l}-\rho_y\right)d\rho_x d\rho_y
\end{aligned}\tag{27}$$

on the property $(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D}) = (\mathbf{AC})\otimes(\mathbf{BD})$, and $S_{c_l}^{\mathrm{v,E}}\left(\rho_x,\rho_y\right) = \frac{\beta_{c_l}K_{c_l}}{K_{c_l}+1}\delta\left(\rho_x-\rho_{c_l,x}\right)\delta\left(\rho_y-\rho_{c_l,y}\right) + \frac{\beta_{c_l}}{K_{c_l}+1}S_{c_l}^{\mathrm{v}}\left(\rho_x,\rho_y\right)$ denotes the effective PAS considering the large-scale fading and LoS channel component.

For the commonly used PAS models, $S_{c_l}^{\mathrm{v}}\left(\rho_x,\rho_y\right)$ achieves the maximum at $\{\rho_x,\rho_y\}=\{\rho_{c_l,x},\rho_{c_l,y}\}$ [21], where $\rho_{c_l,x} = \frac{d_x}{\lambda}\cos\left(\theta_{c_l}\right)\sin\left(\varphi_{c_l}\right)$ and $\rho_{c_l,y} = \frac{d_x}{\lambda}\cos\left(\theta_{c_l}\right)\cos\left(\varphi_{c_l}\right)$ denote the virtual AoDs corresponding to the UEs' real AoDs in horizontal and vertical directions. Moreover, $S_{c_l}^{\mathrm{v}}\left(\rho_x,\rho_y\right)$ decreases with the increasing of $|\rho_x-\rho_{c_l,x}|$ and $|\rho_y-\rho_{c_l,y}|$. Since $\mathrm{asinc}_N^2(x)$ has sharp mainlobe around $x=0$ and very small sidelobes for large N (acts similar as a dirac delta function) [34], the solution of (26) should ensure that $S_{c_l}^{\mathrm{v,E}}\left(\rho_x,\rho_y\right)$ and $\mathrm{asinc}_{N_x}^2\left(\omega_x-\rho_x\right)\mathrm{asinc}_{N_y}^2\left(\omega_y-\rho_y\right)$ reach their maximums at the same $\{\rho_x,\rho_y\}$. Note that the squred aliased sinc function $\mathrm{asinc}_N^2(x)$ achieves its maximum at $x=0$, thus we have

$$\begin{aligned}
\omega_{x,l} &= \rho_{c_l,x} = \frac{d_x}{\lambda}\cos\left(\theta_{c_l}\right)\sin\left(\varphi_{c_l}\right),\\
\omega_{y,l} &= \rho_{c_l,y} = \frac{d_y}{\lambda}\cos\left(\theta_{c_l}\right)\cos\left(\varphi_{c_l}\right).
\end{aligned}\tag{28}$$

With the solution in (28), we can express $\left|\mathbf{p}_{c,l}^{H}\mathbf{p}_{c,m}\right|$ ($l,m\in\{1,2,\cdots,L_c\}$ and $l\neq m$) as

$$\begin{aligned}
&\left|\mathbf{p}_{c,l}^{H}\mathbf{p}_{c,m}\right|\\
&= \left|\left(\tilde{\mathbf{g}}_{N_x}\left(\rho_{c_l,x}\right)\otimes\tilde{\mathbf{g}}_{N_y}\left(\rho_{c_l,y}\right)\right)^{H}\tilde{\mathbf{g}}_{N_x}\left(\rho_{c_m,x}\right)\otimes\tilde{\mathbf{g}}_{N_y}\left(\rho_{c_m,y}\right)\right|\\
&= \left|\left(\tilde{\mathbf{g}}_{N_x}^{H}\left(\rho_{c_l,x}\right)\tilde{\mathbf{g}}_{N_x}\left(\rho_{c_m,x}\right)\right)\otimes\left(\tilde{\mathbf{g}}_{N_y}^{H}\left(\rho_{c_l,y}\right)\tilde{\mathbf{g}}_{N_y}\left(\rho_{c_m,y}\right)\right)\right|\\
&= \left|\sum_{i=0}^{N_x-1}\exp\left(\jmath 2\pi i\left(\rho_{c_m,x}-\rho_{c_l,x}\right)\right)\right.\\
&\qquad\left.\times\sum_{i=0}^{N_x-1}\exp\left(\jmath 2\pi i\left(\rho_{c_m,y}-\rho_{c_l,y}\right)\right)\right|\\
&= \left|\mathrm{asinc}\left(\rho_{c_m,x}-\rho_{c_l,x}\right)\mathrm{asinc}\left(\rho_{c_m,x}-\rho_{c_l,x}\right)\right|\\
&\le \min\left\{\frac{1}{N_x\left(\rho_{c_m,x}-\rho_{c_l,x}\right)},1\right\}\min\left\{\frac{1}{N_y\left(\rho_{c_m,y}-\rho_{c_l,y}\right)},1\right\}.
\end{aligned}\tag{29}$$

The last step is because $|\mathrm{asinc}_N(x)|\le 1$ and the equality holds only when $x=0$. Moreover, since $\mathrm{asinc}_N(x)$ converges to standard sinc function $\mathrm{sinc}(Nx)$ for large N [34] and $\sin x\le 1$, we have $|\mathrm{asinc}_N(x)|\le\frac{1}{Nx}$. From (29), we can see that $\left|\mathbf{p}_{c,l}^{H}\mathbf{p}_{c,m}\right|\overset{N_x,N_y\to\infty}{=}0$ as long as $\rho_{c_m,x}-\rho_{c_l,x}\neq 0$ or $\rho_{c_m,y}-\rho_{c_l,y}\neq 0$ which occurs with probability 1. As a result, the second constraint of (24) is satisfied asymptotically.

For the general case with $b_c > L_c$, we present a heuristic algorithm based on the solution developed in the above. Our strategy is as follows. First, we compute the first L_c columns of $\mathbf{P}_c$ using (28), since transmitting signals in these directions is expected to provide large power gain for the UEs within the cluster. Let $R_{c,x} = \bigcup_{l=1}^{L_c}\left[\rho_{c_l,x}^{\min},\rho_{c,x}^{\max}\right]$ and $R_{c,y} = \bigcup_{l=1}^{L_c}\left[\rho_{c_l,y}^{\min},\rho_{c,y}^{\max}\right]$, we then update the feasible regions of ω_{x,L_c+1} and ω_{y,L_c+1} as $R'_{c,x} = R_{c,x}/\bigcup_{l=1}^{L_c}\left(\omega_{x,l}-N_x^{-1},\omega_{x,l}+N_x^{-1}\right)$ and $R'_{c,y} = R_{c,y}$ (or $R'_{c,x} = R_{c,x}$ and $R'_{c,y} = R_{c,y}/\bigcup_{l=1}^{L_c}\left(\omega_{y,l}-N_y^{-1},\omega_{y,l}+N_y^{-1}\right)$). This operation is to introduce a small gap between ω_{x,L_c+1} and $\omega_{x,l}$ (or ω_{y,L_c+1} and $\omega_{y,l}$) for $l\le L_c$, so that $|\mathbf{p}_{c,l}^{H}\mathbf{p}_{c,L_c+1}|$ can converge to zero asymptotically. Then we perform 2D search to find ω_{x,L_c+1} and ω_{y,L_c+1} within the new feasible regions to maximize $\sum_{l=1}^{L_c}\mathrm{tr}\left(\mathbf{p}_{c,L_c+1}^{H}\mathbf{C}_{c_l}\mathbf{p}_{c,L_c+1}\right)$. Repeating this process, we can find $\omega_{x,i}$ and $\omega_{y,i}$ for $i=L_c+2,\cdots,b_c$ sequentially.

Remark 1: When the number of RF chains at AP is reduced, the effective channel matrix $\tilde{\mathbf{H}}_c$ can be efficiently estimated using the methods in [11] and [32]. No change is needed on the transceiver structure of AP.

C. RELATION BETWEEN AP'S ALTITUDE AND NUMBER OF RF CHAINS

As mentioned in the above, the number of required RF chains on AP is highly correlated with the choice of b_c, i.e., the dimension of the small MIMO system. In particular, when the reduced complexity precoding is employed, the total number of required RF chains is $\sum_{c=1}^{C} b_c$. According to the algorithm in the last subsection, when select the ith ($L_c < i \leq b_c$) first-layer precoding vector, we require $|\omega_{x,i} - \omega_{x,i'}| \geq \frac{1}{N_x}$ or $|\omega_{y,i} - \omega_{y,i'}| \geq \frac{1}{N_y}$, for all $i' < i$. This indicates that b_c is upper bounded by

$$b_c \leq \sum_{l=1}^{L_c} \left(N_x \left(\rho_{c_l,x}^{\max} - \rho_{c_l,x}^{\min}\right)+1\right)\left(N_y \left(\rho_{c_l,y}^{\max} - \rho_{c_l,y}^{\min}\right)+1\right). \tag{30}$$

Without loss of generality, we assume that $\left[\varphi_{c_l} - \Delta\varphi_{c_l}, \varphi_{c_l} + \Delta\varphi_{c_l}\right] \subset \left[0, \frac{\pi}{2}\right]$, $\forall l \in \{1, 2, \cdots, L_c\}$. In this case, (30) can be rewritten as

$$\begin{aligned} b_c \leq \sum_{l=1}^{L_c} \Bigg(& \frac{d_x N_x}{\lambda} \left(\cos\left(\theta_{c_l} - \Delta\theta_{c_l}\right)\sin\left(\varphi_{c_l} + \Delta\varphi_{c_l}\right)\right. \\ & \left. - \cos\left(\theta_{c_l} + \Delta\theta_{c_l}\right)\sin\left(\varphi_{c_l} - \Delta\varphi_{c_l}\right)\right) + 1\Bigg) \\ & \times \left(\frac{d_y N_y}{\lambda} \left(\cos\left(\theta_{c_l} - \Delta\theta_{c_l}\right)\cos\left(\varphi_{c_l} - \Delta\varphi_{c_l}\right)\right.\right. \\ & \left.\left. - \cos\left(\theta_{c_l} + \Delta\theta_{c_l}\right)\cos\left(\varphi_l + \Delta\varphi_{c_l}\right)\right) + 1\right). \end{aligned} \tag{31}$$

Considering the angular spread model in [11], we can express $\Delta\varphi_{c_l}$ and $\Delta\theta_{c_l}$ as

$$\begin{aligned} \Delta\varphi_{c_l} &= \arctan\left(\frac{r}{d_{c_l}}\right), \\ \Delta\theta_{c_l} &= \frac{1}{2}\left(\arctan\left(\frac{d_{c_l} + r}{h}\right) - \arctan\left(\frac{d_{c_l} - r}{h}\right)\right). \end{aligned} \tag{32}$$

where r denotes the scattering radius, d_{c_l} denotes the distance between AP and UE c_l in the $x-y$ plane and h denotes the altitude of the AP. Since AP with higher altitude is expected to have larger coverage area [21], to get a clear insight on the effect AP's altitude, we assume that $d_{c_l} = h$ in the following.

Low Altitude AP: When the altitude of AP is low so that the scattering radius is comparable with h, the angular spread in the vertical direction $\Delta\theta_{c_l}$ is large and it is possible that $N_y \left(\rho_{c_l,y}^{\max} - \rho_{c_l,y}^{\min}\right) > 1$. In this case, we can choose $b_c > L_c$ to improve the power gain of the first-layer precoding at cost of higher hardware complexity.

High Altitude AP: When the altitude of AP is high so that the scattering radius $r \ll h$ or $\frac{r}{h} \to 0$, we have

$$\begin{aligned} \Delta\varphi_{c_l} &= \left.\frac{\mathrm{d}\arctan(x)}{\mathrm{d}x}\right|_{x=\frac{r}{h}} \frac{r}{h} \approx \frac{r}{h}, \\ \Delta\theta_{c_l} &= \frac{1}{2}\left.\frac{\mathrm{d}\arctan(x)}{\mathrm{d}x}\right|_{x=\frac{h+r}{h}} \frac{2r}{h} \approx \frac{r}{h}. \end{aligned} \tag{33}$$

Substituting (33) into (31) and using the results $\sin\Delta\varphi_{c_l} \approx \frac{r}{h}$, $\sin\Delta\theta_{c_l} \approx \frac{r}{h}$, $\cos\Delta\varphi_{c_l} \approx 1$, $\cos\Delta\theta_{c_l} \approx 1$ (since $\frac{r}{h} \to 0$), we can rewrite (31) as

$$\begin{aligned} d_c \leq \sum_{l=1}^{L_c} & \left(\frac{d_x N_x}{\lambda}\frac{2r}{h}\cos\left(\theta_{c_l} - \varphi_{c_l}\right) + 1\right) \\ & \times \left(\frac{d_y N_y}{\lambda}\frac{2r}{h}\sin\left(\theta_{c_l} + \varphi_{c_l}\right) + 1\right). \end{aligned} \tag{34}$$

From (34), we can see that the upper bound decreases with the AP's altitude and increases with the number of AP antennas. In real system, the number of AP antennas cannot be too large, thus a practical selection for high altitude AP is $b_c = L_c$. As a concrete example, consider the recent popular stratospheric AP ($h \approx 20$ km) with a 25×25 rectangular antenna array. Since the scattering radius is commonly less than 100 m [35], we have $\frac{2rN_x}{h} \ll 1$. Thus, it is enough to set $b_c = L_c$.

V. EFFECT OF AoD UNCERTAINTY

As shown in section IV-B, precise AoDs of UEs are needed to compute the first-layer precoding matrix. However, there always exists uncertainty on AoD information due to estimation error and relative movement between AP and UEs, which causes loss on power gain. In practice, fast evaluation of power loss due to AoD uncertainty is essential for link budget and design of AoD estimation scheme. To this end, in this section we propose a new analytic method to estimate the power gain of first-layer precoding with imperfect AoD information.

According to (13), the power gain of first-layer precoding for cluster c can be expressed as

$$\begin{aligned} & \mathcal{G}_c\left(\left\{\omega_{x,i}, \omega_{y,i}\right\}_{i=1}^{b_c}\right) \\ &= \sum_{l=1}^{L_c} \mathrm{tr}\left(\mathbf{P}_c^H \mathbf{C}_{c_l} \mathbf{P}_c\right) \\ &= \sum_{l=1}^{L_c}\sum_{i=1}^{b_c} \mathbf{p}_{c,i}^H \mathbf{C}_{c_l}^{\mathrm{LoS}} \mathbf{p}_{c,i} + \sum_{l=1}^{L_c}\sum_{i=1}^{b_c} \mathbf{p}_{c,i}^H \mathbf{C}_{c_l}^{\mathrm{NLoS}} \mathbf{p}_{c,i} \\ &\triangleq \mathcal{G}_c^{\mathrm{LoS}}\left(\left\{\omega_{x,i}, \omega_{y,i}\right\}_{i=1}^{b_c}\right) + \mathcal{G}_c^{\mathrm{NLoS}}\left(\left\{\omega_{x,i}, \omega_{y,i}\right\}_{i=1}^{b_c}\right). \end{aligned} \tag{35}$$

The first term of right-hand side of (35) indicates the power gain on LoS channel component and the second term indicates the power gain on NLoS channel component. Let $\left\{\hat{\omega}_{x,i}, \hat{\omega}_{y,i}\right\}_{i=1}^{b_c}$ denote the estimates of $\left\{\omega_{x,i}, \omega_{y,i}\right\}_{i=1}^{b_c}$ computed based on the imperfect AoD information, the percentage of loss on power gain can be expressed as

$$\eta_c = 1 - \frac{\mathcal{G}_c\left(\left\{\hat{\omega}_{x,i}, \hat{\omega}_{y,i}\right\}_{i=1}^{b_c}\right)}{\mathcal{G}_c\left(\left\{\omega_{x,i}, \omega_{y,i}\right\}_{i=1}^{b_c}\right)}. \tag{36}$$

From (36), to investigate the impact of AoD uncertainty, the most important is to derive the analytical expressions of $\mathcal{G}_c^{\mathrm{LoS}}(\left\{\hat{\omega}_{x,i}, \hat{\omega}_{y,i}\right\}_{i=1}^{b_c})$ and $\mathcal{G}_c^{\mathrm{NLoS}}(\left\{\hat{\omega}_{x,i}, \hat{\omega}_{y,i}\right\}_{i=1}^{b_c})$.

A. POWER GAIN ON LoS CHANNEL COMPONENT

Based on the definition of $\mathbf{C}_{c_l}^{\mathrm{LoS}}$ in (6) and using a similar procedure as that in (27), we can express $\mathcal{G}_c^{\mathrm{LoS}}(\{\hat{\omega}_{x,i}, \hat{\omega}_{y,i}\}_{i=1}^{b_c})$ as

$$\begin{aligned}\mathcal{G}_c^{\mathrm{LoS}}&\left(\{\hat{\omega}_{x,i}, \hat{\omega}_{y,i}\}_{i=1}^{b_c}\right)\\ &= \sum_{l=1}^{L_c}\sum_{i=1}^{b_c}\frac{\beta_{c_l}K_{c_l}}{K_{c_l}+1}N_xN_y\\ &\quad\times \mathrm{asinc}_{N_x}^2\left(\hat{\omega}_{x,i}-\rho_{c_l,x}\right)\mathrm{asinc}_{N_y}^2\left(\hat{\omega}_{y,i}-\rho_{c_l,y}\right).\end{aligned} \tag{37}$$

The above expression can be used to compute the power gain on the LoS channel with arbitrary estimated $\{\hat{\omega}_{x,i}, \hat{\omega}_{y,i}\}_{i=1}^{b_c}$. To get more insight, we consider the special case with $b_c = L_c = 1$. In this case, using the property that $\mathrm{asinc}_N(x)$ converges to $\mathrm{sinc}(Nx)$ for large N [34], we have

$$\mathcal{G}_c^{\mathrm{LoS}}\left(\{\hat{\omega}_{x,i}, \hat{\omega}_{y,i}\}_{i=1}^{b_c}\right) \approx \frac{\beta_{c_1}K_{c_1}}{K_{c_1}+1}\frac{\sin^2\left(N_x\delta_x\right)}{N_x\pi^2\delta_x^2}\frac{\sin^2\left(N_y\delta_y\right)}{N_x\pi^2\delta_y^2}, \tag{38}$$

where $\delta_x = \hat{\rho}_{c_1,x} - \rho_{c_1,x}$, $\delta_y = \hat{\rho}_{c_1,y} - \rho_{c_1,y}$ with $\{\hat{\rho}_{c_l,x}, \hat{\rho}_{c_l,y}\}$ denoting the estimates of $\{\rho_{c_l,x}, \rho_{c_l,y}\}$. Note that when the AoD estimation is perfect, we have $\delta_x = \delta_y = 0$. From (38), we can see that the power gain on LoS channel component becomes sensitive to the AoD error when N_x and N_y increase. Moreover, when the AoD error is large so that $\delta_x^2 \gg N_x^{-1}$ and $\delta_y^2 \gg N_y^{-1}$, the power gain diminishes to zero. In this case, almost no power is delivered by the LoS channel.

B. POWER GAIN ON NLoS COMPONENT

Similarly, substituting (7) into (35), the power gain on NLoS channel component can be obtained as

$$\begin{aligned}\mathcal{G}_c^{\mathrm{NLoS}}&\left(\{\hat{\omega}_{x,i}, \hat{\omega}_{y,i}\}_{i=1}^{b_c}\right)\\ &= \sum_{l=1}^{L_c}\sum_{i=1}^{b_c}\frac{\beta_{c_l}N_xN_y}{K_{c_l}+1}\int_{\rho_{c_l,x}^{\min}}^{\rho_{c_l,x}^{\max}}\int_{\rho_{c_l,y}^{\min}}^{\rho_{c_l,y}^{\max}} S_{c_l}^{\mathrm{v}}\left(\rho_x,\rho_y\right)\\ &\quad\times \mathrm{asinc}_{N_x}^2\left(\hat{\omega}_{x,i}-\rho_x\right)\mathrm{asinc}_{N_y}^2\left(\hat{\omega}_{y,i}-\rho_y\right)\mathrm{d}\rho_x\mathrm{d}\rho_y.\end{aligned} \tag{39}$$

The above integrals are difficult to solve in closed-form since the integrand involves aliased sinc function and the expression of $S_{c_l}^{\mathrm{v}}\left(\rho_x,\rho_y\right)$ is complex for most of the existing PAS models [21], [35]. Thus, directly evaluating the power gain using (39) is challenging. In the following, we propose a simple method to estimate the power gain with arbitrary AoD error and PAS model. The key idea is to approximate the integrand of (39) using some functions that have simple structures but capture most of the properties of original integrand.

As mentioned in section IV-B, for the commonly used PAS model, $S_{c_l}^{\mathrm{v}}\left(\rho_x,\rho_y\right)$ achieves the maximum at $\rho_x = \rho_{c_l,x} = \frac{d_x}{\lambda}\cos\left(\theta_{c_l}\right)\sin\left(\varphi_{c_l}\right)$ and $\rho_y = \rho_{c_l,y} = \frac{d_x}{\lambda}\cos\left(\theta_{c_l}\right)\cos\left(\varphi_{c_l}\right)$, and decreases with the increasing of $|\rho_x - \rho_{c_l,x}|$ and $|\rho_y - \rho_{c_l,y}|$. This motivates us to approximate $S_{c_l}^{\mathrm{v}}\left(\rho_x,\rho_y\right)$ using the following simple function

$$\begin{aligned}\hat{S}_{c_l}^{\mathrm{v}}\left(\rho_x,\rho_y\right) = S_{c_l}^{\mathrm{v}}\left(\rho_{c_l,x},\rho_{c_l,y}\right) &- A_x\left|\rho_x-\rho_{c_l,x}\right|\\ &- A_y\left|\rho_y-\rho_{c_l,y}\right|,\end{aligned} \tag{40}$$

where $A_x, A_y > 0$ can be interpreted as the decreasing rates of $\hat{S}_{c_l}^{\mathrm{v}}\left(\rho_x,\rho_y\right)$ with respect to $\left|\rho_x-\rho_{c_l,x}\right|$ and $\left|\rho_y-\rho_{c_l,y}\right|$, which are determined by solving the following problem.

$$\begin{aligned}\min_{A_x,A_y>0}\ &\int_{\rho_{c_l,x}^{\min}}^{\rho_{c_l,x}^{\max}}\int_{\rho_{c_l,y}^{\min}}^{\rho_{c_l,y}^{\max}}\left(\hat{S}_{c_l}^{\mathrm{v}}\left(\rho_x,\rho_y\right)-S_{c_l}^{\mathrm{v}}\left(\rho_x,\rho_y\right)\right)^2\mathrm{d}\rho_x\mathrm{d}\rho_y.\\ \text{s.t. }&\hat{S}_{c_l}^{\mathrm{v}}\left(\rho_x,\rho_y\right)\geq 0,\\ &\forall\rho_x\in\left[\rho_{c_l,x}^{\min},\rho_{c_l,x}^{\max}\right],\quad \forall\rho_y\in\left[\rho_{c_l,y}^{\min},\rho_{c_l,y}^{\max}\right].\end{aligned} \tag{41}$$

Inserting (40) into (41) and using some algebraic manipulations, we can rewrite the problem as

$$\begin{aligned}\min_{A_x,A_y>0}\ &t_1A_x^2+t_2A_y^2-t_3A_x-t_4A_y,\\ \text{s.t. }&u_1A_x+u_2A_y-S_{c_l}^{\mathrm{v}}\left(\rho_{c_l,x},\rho_{c_l,y}\right)\leq 0,\end{aligned} \tag{42}$$

where t_1, t_2, t_3, t_4 and u_1, u_2 are given by

$$\begin{aligned}t_1 &= \frac{\left(\rho_{c_l,x}^{\max}-\rho_{c_l,x}\right)^3+\left(\rho_{c_l,x}-\rho_{c_l,x}^{\min}\right)^3}{3}\left(\rho_{c_l,y}^{\max}-\rho_{c_l,y}^{\min}\right),\\ t_2 &= \frac{\left(\rho_{c_l,y}^{\max}-\rho_{c_l,y}\right)^3+\left(\rho_{c_l,y}-\rho_{c_l,y}^{\min}\right)^3}{3}\left(\rho_{c_l,x}^{\max}-\rho_{c_l,x}^{\min}\right),\\ t_3 &= \int_{\rho_{c_l,x}^{\min}}^{\rho_{c_l,x}^{\max}}\int_{\rho_{c_l,y}^{\min}}^{\rho_{c_l,y}^{\max}}2\left|\rho_x-\rho_{c_l,x}\right|,\\ &\quad\times\left(S_{c_l}^{\mathrm{v}}\left(\rho_{c_l,x},\rho_{c_l,y}\right)-S_{c_l}^{\mathrm{v}}\left(\rho_x,\rho_y\right)\right)\mathrm{d}\rho_x\mathrm{d}\rho_y,\\ t_4 &= \int_{\rho_{c_l,x}^{\min}}^{\rho_{c_l,x}^{\max}}\int_{\rho_{c_l,y}^{\min}}^{\rho_{c_l,y}^{\max}}2\left|\rho_y-\rho_{c_l,y}\right|,\\ &\quad\times\left(S_{c_l}^{\mathrm{v}}\left(\rho_{c_l,x},\rho_{c_l,y}\right)-S_{c_l}^{\mathrm{v}}\left(\rho_x,\rho_y\right)\right)\mathrm{d}\rho_x\mathrm{d}\rho_y,\\ u_1 &= \max\left\{\rho_{c_l,x}^{\max}-\rho_{c_l,x},\rho_{c_l,x}-\rho_{c_l,x}^{\min}\right\},\\ u_2 &= \max\left\{\rho_{c_l,y}^{\max}-\rho_{c_l,y},\rho_{c_l,y}-\rho_{c_l,y}^{\min}\right\}.\end{aligned} \tag{43}$$

By exploiting the Karush-Kuhn-Tucker condition, the solution of (42) can be obtained as

$$\begin{aligned}A_x &= \frac{t_3}{2t_1}-\frac{u_1}{2t_1}\left(\frac{t_3u_1}{2t_1}+\frac{t_4u_2}{2t_2}-S_{c_l}^{\mathrm{v}}\left(\rho_{c_l,x},\rho_{c_l,y}\right)\right)\\ &\quad\times\left(\frac{u_1^2}{2t_1}+\frac{u_2^2}{2t_2}\right)^{-1},\\ A_y &= \frac{t_4}{2t_2}-\frac{u_2}{2t_2}\left(\frac{t_3u_1}{2t_1}+\frac{t_4u_2}{2t_2}-S_{c_l}^{\mathrm{v}}\left(\rho_{c_l,x},\rho_{c_l,y}\right)\right)\\ &\quad\times\left(\frac{u_1^2}{2t_1}+\frac{u_2^2}{2t_2}\right)^{-1}.\end{aligned} \tag{44}$$

Then we consider the approximation of squared aliased sinc function. Again with the large N property of $\mathrm{asinc}_N(x)$,

we have

$$\text{asinc}^2_{N_j}\left(\hat{\omega}_{j,i}-\rho_j\right)\approx\begin{cases}\dfrac{\sin^2\left(N_j\pi\left(\hat{\omega}_{j,i}-\rho_j\right)\right)}{N_j^2\pi^2\left(\hat{\omega}_{x,i}-\rho_j\right)^2}, \\ \qquad \left|\hat{\omega}_{j,i}-\rho_j\right|\le N_j^{-1}, \\ 0, \quad \left|\hat{\omega}_{j,i}-\rho_j\right|>N_j^{-1},\end{cases} \tag{45}$$

where $j\in\{x,y\}$. Note that in (45) we have assume that the squared aliased sinc function has zero value outside its mainlobe since the sidelobes are very small for large N_j [34]. According to [36], $\sin\pi x$ can be precisely approximated by $\sin\pi x=\pi x\frac{q_1-q_2x^2+x^4}{q_1+q_3x^2}$ at $x\in[-1,1]$ with $q_1=0.9\pi$, $q_2=3.83$ and $q_3=0.27\pi$. Thus, (45) can be further rewritten as

$$\text{asinc}^2_{N_j}\left(\hat{\omega}_{j,i}-\rho_j\right)\approx\begin{cases}\left(\dfrac{q_1-q_2N_j^2\left(\hat{\omega}_{j,i}-\rho_j\right)^2+N_j^4\left(\hat{\omega}_{j,i}-\rho_j\right)^4}{q_1+q_3N_j^2\left(\hat{\omega}_{j,i}-\rho_j\right)^2}\right)^2, & \left|\hat{\omega}_{j,i}-\rho_j\right|\le N_j^{-1} \\ 0, \quad \left|\hat{\omega}_{j,i}-\rho_j\right|>N_j^{-1}.\end{cases} \tag{46}$$

Using the above approximations, the closed-form expression of power gain on NLoS channel component can be obtained by substituting (40) and (46) into (39) and solving the integrals directly, which is given in the below. A brief derivation is presented in the Appendix.

$$\begin{aligned}&\mathcal{G}_c^{\text{NLoS}}\left(\left\{\omega_{x,i},\omega_{y,i}\right\}_{i=1}^{b_c}\right)\\&\approx\sum_{l=1}^{L_c}\sum_{i=1}^{b_c}\frac{\beta_{c_l}N_xN_y}{K_{c_l}+1}\\&\quad\times\left\{S_{c_l}^{\text{v}}\left(\rho_{c_l,x},\rho_{c_l,y}\right)\mathcal{I}_{i,x}\left(\rho_{c_l,x}^{\min},\rho_{c_l,x}^{\max}\right)\mathcal{I}_{i,y}\left(\rho_{c_l,y}^{\min},\rho_{c_l,y}^{\max}\right)\right.\\&\quad-A_x\left(\mathcal{J}_{i,x}\left(\rho_{c_l,x},\rho_{c_l,x}^{\max}\right)-\mathcal{J}_{i,x}\left(\rho_{c_l,x}^{\min},\rho_{c_l,x}\right)\right)\mathcal{I}_{i,y}\\&\quad\times\left(\rho_{c_l,y}^{\min},\rho_{c_l,y}^{\max}\right)-A_y\mathcal{I}_{i,x}\left(\rho_{c_l,x}^{\min},\rho_{c_l,x}^{\max}\right)\\&\quad\left.\times\left(\mathcal{J}_{i,y}\left(\rho_{c_l,y},\rho_{c_l,y}^{\max}\right)-\mathcal{J}_{i,y}\left(\rho_{c_l,y}^{\min},\rho_{c_l,y}\right)\right)\right.\end{aligned} \tag{47}$$

Wherein, the functions $\mathcal{I}_{i,j}(a,b)$ and $\mathcal{J}_{i,j}(a,b)$, $j\in\{x,y\}$, are given by

$$\begin{aligned}&\mathcal{I}_{i,j}(a,b)\\&=\frac{\phi(1)-\gamma(1)}{N_j}-2q_2\frac{\phi(3)-\gamma(3)}{N_jq_1}\\&\quad+\left(2q_1+q_2^2\right)\frac{\phi(5)-\gamma(5)}{N_jq_1^2}-2q_2\frac{\phi(7)-\gamma(7)}{N_jq_1^2}\\&\quad+\frac{\phi(9)-\gamma(9)}{N_jq_1^2},\end{aligned}$$

$$\begin{aligned}&\mathcal{J}_{i,j}(a,b)\\&=\left(\hat{\omega}_{j,i}-\rho_{c_l,j}\right)\mathcal{I}_{i,j}(a,b)+\frac{|\phi(3)|-|\gamma(3)|}{N_j^2}\\&\quad-2q_2\frac{|\phi(5)|-|\gamma(5)|}{N_j^2q_1}+\left(2q_1+q_2^2\right)\frac{|\phi(7)|-|\gamma(7)|}{N_j^2q_1^2}\\&\quad-2q_2\frac{|\phi(9)|-|\gamma(9)|}{N_j^2q_1^2}+\frac{|\phi(11)|-|\gamma(11)|}{N_j^2q_1^2},\end{aligned}$$

$$\begin{aligned}&\phi(u)\\&=\frac{\text{sgn}\left(b-\hat{\omega}_{j,i}\right)}{u}\left|\min\left\{N_j\left(b-\hat{\omega}_{j,i}\right),1\right\}\right|^u\\&\quad\times{}_2F_1\left(2,\frac{u}{2},\frac{u}{2}+1,-\frac{\left|\min\left\{N_j\left(b-\omega_{j,i}\right),1\right\}\right|}{q_1q_3^{-1}}\right),\\&\gamma(u)\\&=\frac{\text{sgn}\left(a-\hat{\omega}_{j,i}\right)}{u}\left|\max\left\{N_j\left(a-\hat{\omega}_{j,i}\right),-1\right\}\right|^u\\&\quad\times{}_2F_1\left(2,\frac{u}{2},\frac{u}{2}+1,-\frac{\left|\max\left\{N_j\left(a-\omega_{j,i}\right),-1\right\}\right|}{q_1q_3^{-1}}\right).\end{aligned} \tag{48}$$

In practical implementation, to fast estimate the power gain, we can build off-line table of A_x and A_y for different AoDs and angular spreads. Thus, no integral computation is needed.

Then let us consider the (l,i)th term in the summation of (47) (denoted by $\mathcal{G}_{c,l,i}^{\text{NLoS}}$ for convenience). From (35), $\mathcal{G}_{c,l,i}^{\text{NLoS}}$ can be viewed as the power gain of ith first-layer precoding vector (i.e., the ith column of $\mathbf{P}_c$) on the NLoS channel component of UE c_l. When $\hat{\omega}_{x,i}=\rho_{c_l,x}$ and $\hat{\omega}_{y,i}=\rho_{c_l,y}$, $\mathcal{G}_{c,l,i}^{\text{NLoS}}$ achieves its maximum according to section IV-B. When $\hat{\omega}_{j,i}\notin\left[\rho_{c_l,j}^{\min},\rho_{c_l,j}^{\max}\right]$ ($j\in\{x,y\}$), for large N_j, it can be verified that $\left|\min\left\{N_j\left(b-\omega_{j,i}\right),1\right\}\right|=\left|\max\left\{N_j\left(a-\omega_{j,i}\right),-1\right\}\right|=1$ and $\text{sgn}\left(a-\hat{\omega}_{j,i}\right)=\text{sgn}\left(b-\hat{\omega}_{j,i}\right)$, $\forall a,b\in\{\rho_{c_l,j}^{\min},\rho_{c_l,j}^{\max},\rho_{c_l,j}\}$. In this case, $\mathcal{I}_{i,j}(a,b)=\mathcal{J}_{i,j}(a,b)=\mathcal{G}_{c,l,i}^{\text{NLoS}}=0$, and almost no power gain is achieved. Imaging that $\hat{\omega}_{j,i}$ is originally designed in $\left[\rho_{c_l,j}^{\min},\rho_{c_l,j}^{\max}\right]$, but falls outside $\left[\rho_{c_l,j}^{\min},\rho_{c_l,j}^{\max}\right]$ due to the AoD errors. No power gain will be obtained on UE c_l according to the analysis in the above. However, since we have multiple precoding vectors with different $\hat{\omega}_{j,i}$. It is possible that some other precoding vectors with $\hat{\omega}_{j,i'}$ originally designed not in $\left[\rho_{c_l,j}^{\min},\rho_{c_l,j}^{\max}\right]$ achieve large power gain. This property can potentially be used to design first-layer precoding scheme which is robust to AoD uncertainty. For example, one possible solution is to introduce some redundancy when choosing $\hat{\omega}_{j,i}$. This can be considered in future work.

VI. SIMULATION RESULTS AND DISCUSSION

This section evaluates the performance of the proposed precoding scheme via Matlab simulations. We consider the SE performance, which is defined as $\text{SE}=\mathbb{E}\left[\sum_{c=1}^{C}\sum_{l=1}^{L_c}\log_2\left(1+\text{SINR}_{c_l}\right)\right]$, where SINR_{c_l} denotes the received SINR at UE c_l. According to (11), we have

$$\begin{aligned}&\text{SINR}_{c_l}\\&=\frac{p_{c_l}\left|\mathbf{h}_{c,l}^H\mathbf{P}_c\mathbf{w}_{c,l}\right|^2}{\sum_{j=1,j\neq l}^{L_c}p_{c_j}\left|\mathbf{h}_{c,l}^H\mathbf{P}_c\mathbf{w}_{c,j}\right|^2+\sum_{c'=1,c'\neq c}^{C}\sum_{j=1}^{L_{c'}}p_{c'_j}\left|\mathbf{h}_{c,l}^H\mathbf{P}_{c'}\mathbf{w}_{c',j}\right|^2+\sigma^2}.\end{aligned} \tag{49}$$

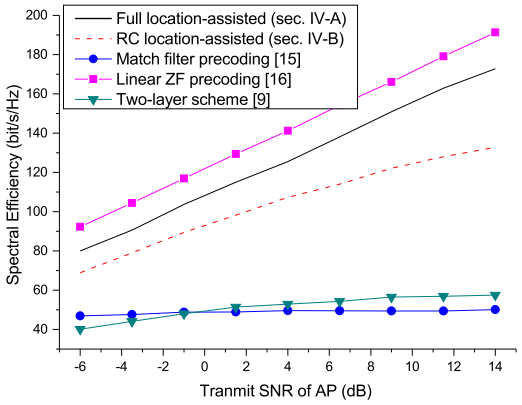


FIGURE 3. Spectral efficiency of proposed precoding scheme as a function of transmit SNR of AP, where $\Delta\varphi_{c_l} = 5°$ and $\Delta\theta_{c_l} = 2.5°$, $b_c = L_c$, $K_{c_l} = 0$ dB, $N_x = N_y = 25$.

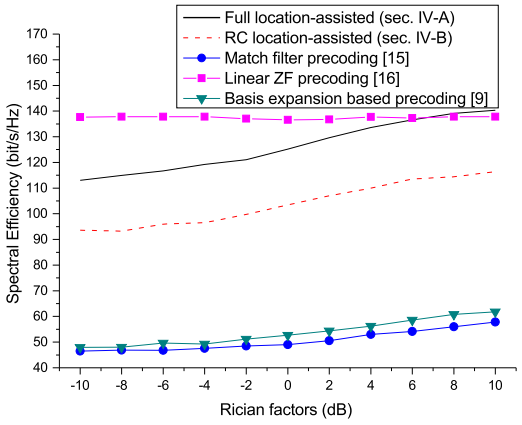


FIGURE 4. Spectral efficiency as a function of Rician factor, where $\Delta\varphi_{c_l} = 5°$ and $\Delta\theta_{c_l} = 2.5°$, $b_c = L_c$, SNR= 3 dB.

We assume perfect power control at the AP so that $\frac{p_{c_l}\beta_{c_l}}{\sigma^2}$ is the same for all UEs, and define $\frac{p_{c_l}\beta_{c_l}}{\sigma^2}$ as the transmit SNR (considering the large-scale fading) of AP. In this way, we do not need to consider the specific large-scale fading model. The PAS of channel is model as $S(\varphi,\theta) = S_h(\varphi)S_v(\theta)$, where $S_h(\varphi)$ and $S_v(\theta)$ denote the PASs in horizontal and vertical directions, respectively. As in [21] and [22], $S_h(\varphi)$ and $S_v(\theta)$ are modeled using the Von Mises distribution and truncated Laplacian distribution, respectively. In all figures, we name the precoding scheme in section IV-A as the *full location-assisted precoding* and name the reduced complexity precoding scheme in section IV-B as *RC location-assisted precoding*.

We first consider the effect of transmit SNR of AP and Rician factor on the SE of proposed precoding schemes in Fig. 3 and Fig. 4. We assume that there are three UE clusters with each containing 5 UEs. The horizontal AoDs of UEs in the three clusters are uniformly distributed in $[-55°,-35°]$, $[-15°,5°]$ and $[30°,50°]$ respectively, and the vertical AoDs of UEs in the three clusters are uniformly distributed in $[60°,70°]$, $[45°,55°]$ and $[30°,40°]$ respectively. The number of AP antennas is set to $N_x = N_y = 25$.

Fig. 3 compares the SEs of proposed precoding schemes with match filter precoding [15], linear ZF precoding [16] and basis expansion based precoding [17]. Note that the number of RF chains required in the match filter precoding and linear ZF precoding is equal to that of the full location-assisted precoding (i.e., $N_x \times N_y$). For the basis expansion based precoding, the required number of RF chains is equal to the number of selected orthogonal basis, which is set to $\sum_{c=1}^{C} b_c$, i.e., same with the RC location-assisted precoding. The detailed description on the choice of number of orthogonal basis can be found in [17]. We can see that the performance gap between RC location-assisted precoding and full location-assisted precoding is less than 1 bit/s/Hz per UE in small SNR region. This makes RC location-assisted precoding attractive when the altitude of AP is very high and the signals suffer from severe large-scale fading. In high SNR region, the performance gap increases because RC location-assisted precoding suffers from performance floor due to the existence of residual ICI. On the other hand, it is seen that the linear ZF precoding achieves the best performance. However, the gain is at the cost of very high hardware/computation complexity and more training resources for channel estimation. This may make it infeasible when applied on AP.

Fig. 4 show the SEs of proposed precoding schemes as a function of Rician factors, where we assume the Rician factors for all UEs are the same, i.e., $K_{c_l} = K$. The transmit SNR of AP is set to 3 dB. From the figure, it is seen that the SE increases as the LoS channel becomes stronger. This is quite different from the small-scale MIMO system, where the LoS component may be harmful since it introduces strong correlation between channels of different UEs [37]. However, this problem is alleviated in the massive MIMO system because the large-scale antenna arrays have better spatial resolution.

Then we consider the effect of AP's altitude on the SE of RC location-assisted precoding. To get a clear insight, we assume that there is one UE cluster and the number of UEs in the cluster is $L_c = 5$. The AoDs of UEs in horizontal and vertical directions are set to $[43.8°, 25.3°, 44.4°, 44.0°, 18.7°]$ and $[51.2°, 41.9°, 55.6°, 58.9°, 51.4°]$, respectively. We consider two cases, i.e., 1) the number of AP's RF chains is $b_c = L_c$ and 2) the number of AP's RF chains is $b_c = 2L_c$. As shown in section IV-C, when altitude of AP is high, we cannot find more than L_c first-layer precoding vectors since the angular spread is too small. In this case, to get the simulation result for $b_c = 2L_c$, we enlarge the feasible regions of $\omega_{x,l}$ and $\omega_{y,l}$ as $[-d_x/\lambda, d_x/\lambda]$ and $[-d_y/\lambda, d_y/\lambda]$, respectively, and find optimal $\omega_{x,l}$ and $\omega_{y,l}$ in the new regions using the algorithm in section IV-B. From Fig. 5, we can see that when the altitude of AP is small, adding more RF chains at AP is beneficial to improve SE. However, as the altitude of AP increases ($h > 3000$ m), adding the number of RF chains almost cannot provide any SE gain.

Then we consider the effect of AoD uncertainty on RC location-assisted precoding. Due to the exist of guard

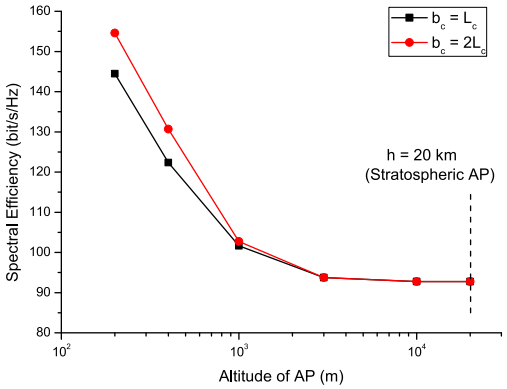


FIGURE 5. Spectral efficiency of RC location-assisted precoding with different AP altitudes, where $K_{c_l} = -5$ dB, SNR = 3 dB.

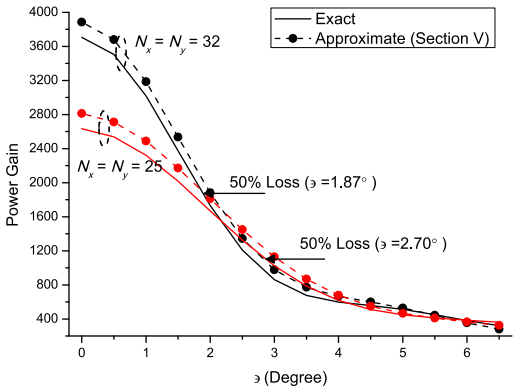


FIGURE 6. Power gain of first-layer precoding with different AoD errors, $K_{c_l} = 0$ dB, $L_c = 5$, $\Delta\varphi_{c_l} = 5°$ and $\Delta\theta_{c_l} = 5°$, $b_c = L_c$.

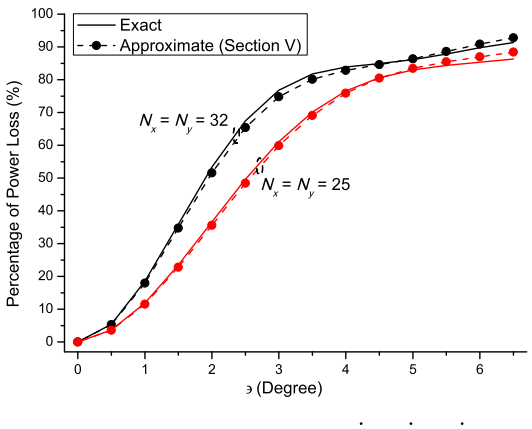


FIGURE 7. Percentage of loss on power gain with different AoD errors, $K_{c_l} = 0$ dB, $L_c = 5$, $\Delta\varphi_{c_l} = 5°$ and $\Delta\theta_{c_l} = 5°$, $b_c = L_c$.

interval between different UE clusters, the AoD uncertainty has little effect on the ICI power for other clusters. Thus, without loss of generality, we consider again the scenario with one UE cluster and let $L_c = 5$. The AoDs of UEs are the same with that in Fig. 5. The AoD errors for horizontal and vertical directions are set to $\epsilon[1, 1, -1, -1, 1]$ and $\epsilon[-1, 1, -1, 1, -1]$, respectively, where $\epsilon > 0$ denotes the absolute value of errors. In Fig. 6 and Fig. 7, we can see that the power gain of first-layer precoding becomes sensitive to AoD errors when the number of AP antennas increases. For example, if $N_x = N_y = 25$, the power gain reduces to 50% of the maximum value (i.e., 3 dB loss) when $\epsilon = 2.70°$. However, this value becomes $\epsilon = 1.87°$ when $N_x = N_y = 32$. Moreover, we can see that the analytical expressions of power gain derived in section V provide good approximation to the real performance.

VII. CONCLUSION

This paper studies the low-complexity precoding for 3D massive MIMO in air-to-ground transmission. By exploiting the slow time-varying parameters, we propose a location-assisted two-layer precoding scheme for downlink transmission. Through proper UE clustering, we show that the first-layer precoding matrix can be approximated using a constant-envelope matrix, which results in significant reduction on required number of RF chains on AP. Since the AoD uncertainty is inevitable in practice, we investigate the effect of AoD uncertainty on the performance of proposed precoding scheme and present a new analytic method to fast estimate the power loss due to AoD errors. Our results show that the location-assisted precoding outperforms match filter and basis expansion precoding in the air-to-ground transmission. Moreover, it is seen that more accurate AoD information is required in precoding when the number of AP antennas increases.

APPENDIX

By substituting (40) and (46) into (39), we can express $\mathcal{G}_c^{\mathrm{NLoS}}\left(\{\omega_{x,i}, \omega_{y,i}\}_{i=1}^{b_c}\right)$ as that in (47), where $\mathcal{I}_{i,j}(a,b)$ and $\mathcal{J}_{i,j}(a,b)$ are given by

$$\mathcal{I}_{i,j}(a,b) = \int_{\max\left\{a-\hat{\omega}_{j,i},-\frac{1}{N_j}\right\}}^{\min\left\{b-\hat{\omega}_{j,i},\frac{1}{N_j}\right\}} \left(\frac{q_1 - q_2 N_j^2\rho_j^2 + N_j^4\rho_j{}^4}{q_1 + q_3 N_j^2\rho_j{}^2}\right)^2 \mathrm{d}\rho_j, \tag{50a}$$

$$\mathcal{J}_{i,j}(a,b) = \int_{\max\left\{a-\hat{\omega}_{j,i},-\frac{1}{N_j}\right\}}^{\min\left\{b-\hat{\omega}_{j,i},\frac{1}{N_j}\right\}} \left(\rho_j - \rho_{c_l,j} + \hat{\omega}_{j,i}\right) \times \left(\frac{q_1 - q_2 N_j^2\rho_j^2 + N_j^4\rho_j{}^4}{q_1 + q_3 N_j^2\rho_j{}^2}\right)^2 \mathrm{d}\rho_j. \tag{50b}$$

The remaining problem is to solve the integrals in (50). In the below, we derive the solution for (50a) and the solution for (50b) can be obtained in the same way. Since the integrand of (50a) is an even function of ρ_j, we can rewrite (50a) as

$$\begin{aligned}\mathcal{I}_{i,j}(a,b) &= \mathrm{sgn}\left(b-\hat{\omega}_{j,i}\right) \times \int_0^{\left|\min\left\{b-\hat{\omega}_{j,i},N_j^{-1}\right\}\right|} \left(\frac{q_1 - q_2 N_j^2\rho_j^2 + N_j^4\rho_j{}^4}{q_1 + q_3 N_j^2\rho_j{}^2}\right)^2 \mathrm{d}\rho_j \\ &\quad - \mathrm{sgn}\left(a-\hat{\omega}_{j,i}\right) \times \int_0^{\left|\max\left\{a-\hat{\omega}_{j,i},-N_j^{-1}\right\}\right|} \left(\frac{q_1 - q_2 N_j^2\rho_j^2 + N_j^4\rho_j{}^4}{q_1 + q_3 N_j^2\rho_j{}^2}\right)^2 \mathrm{d}\rho_j.\end{aligned} \tag{51}$$

By using some algebraic manipulations, the first term of right-hand side of (51) becomes

$$\begin{aligned}
&\operatorname{sgn}\left(b-\hat{\omega}_{j,i}\right)\int_0^{\left|\min\left\{b-\omega_{j,i},N_j^{-1}\right\}\right|}\left(\frac{q_1-q_2N_j^2\rho_j^2+N_j^4\rho_j{}^4}{q_1+q_3N_j^2\rho_j{}^2}\right)^2\mathrm{d}\rho_j\\
&\overset{t=N_j^2\rho_j^2}{=}\frac{\operatorname{sgn}\left(b-\hat{\omega}_{j,i}\right)}{2N_j}\int_0^{N_j^2\left|\min\left\{b,N_j^{-1}\right\}\right|^2}\\
&\quad\times\frac{q_1^2t^{-\frac{1}{2}}-2q_1q_2t^{\frac{1}{2}}+\left(2q_1+q_2^2\right)t^{\frac{3}{2}}-2q_2t^{\frac{5}{2}}+t^{\frac{7}{2}}}{\left(q_1+q_3t\right)^2}\mathrm{d}t\\
&=\frac{\left|\min\left\{N_jb,1\right\}\right|^2}{2N_j}\int_0^1\left(1+\frac{q_3}{q_2}\left|\min\left\{N_jb,1\right\}\right|^2t\right)^{-2}\\
&\quad\times\left(t^{-\frac{1}{2}}-\frac{2q_2}{q_1}\left|\min\{N_jb,1\}\right|t^{\frac{1}{2}}\right.\\
&\quad+\frac{2q_1+q_2^2}{q_1^2}\left|\min\{N_jb,1\}\right|^3t^{\frac{3}{2}}-\frac{2q_2}{q_1^2}\left|\min\left\{N_jb,1\right\}\right|^5t^{\frac{5}{2}}\\
&\quad\left.+\frac{1}{q_1^2}\left|\min\left\{N_jb,1\right\}\right|^7t^{\frac{7}{2}}\right)\mathrm{d}t. \qquad (52)
\end{aligned}$$

By exploiting [19, 9.111], the closed-form expression of (51) can be obtained as

$$\begin{aligned}
&\operatorname{sgn}\left(b-\hat{\omega}_{j,i}\right)\int_0^{\left|\min\left\{b-\omega_{j,i},N_j^{-1}\right\}\right|}\left(\frac{q_1-q_2N_j^2\rho_j^2+N_j^4\rho_j{}^4}{q_1+q_3N_j^2\rho_j{}^2}\right)^2\mathrm{d}\rho_j\\
&=\frac{1}{N_j}\phi\left(1\right)-\frac{2q_2}{N_jq_1}\phi\left(3\right)+\frac{2q_1+q_2^2}{N_jq_1^2}\phi\left(5\right)-\frac{2q_2}{N_jq_1^2}\phi\left(7\right)\\
&\quad+\frac{1}{N_jq_1^2}\phi\left(9\right), \qquad (53)
\end{aligned}$$

where $\phi\left(t,l\right)$ is defined as

$$\begin{aligned}
\phi\left(l\right)=&\frac{\operatorname{sgn}\left(b-\omega_{j,i}\right)}{l}\left|\min\left\{N_j\left(b-\omega_{j,i}\right),1\right\}\right|^l\\
&\times{}_2F_1\left(2,\frac{l}{2},\frac{l}{2}+1,-\frac{\left|\min\left\{N_j\left(b-\omega_{j,i}\right),1\right\}\right|}{q_1q_3^{-1}}\right). \qquad (54)
\end{aligned}$$

Using the similar approach, we can derive the closed-form expression for the second term of right-hand side of (51). This completes the proof.

REFERENCES

[1] *Project LOON*, accessed on 2013. [Online]. Available: https://www.google.com/loon/

[2] *Connecting the World From the Sky*, accessed on 2014. [Online]. Available: https://internet.org/projects

[3] "5G concept," IMT-2020 (5G) promotion group, White Paper, Feb. 2015. [Online]. Available: http://www.imt-2020.cn/en

[4] A. Al-Hourani and S. Kandeepan, "Cognitive relay nodes for airborne LTE emergency networks," in *Proc. 7th Int. Conf. Signal Process. Commun. Syst. (ICSPCS)*, Gold Coast, QLD, Australia, Dec. 2013, pp. 1–9.

[5] "The role of deployable aerial communications architecture in emergency communications and recommended next steps," Federal Communications Commission, Washington, DC, USA, White Paper, 2009. [Online]. Available: http://hraunfoss.fcc.gov/edocspublic/attachmatch

[6] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless communications with unmanned aerial vehicles: Opportunities and challenges," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 36–42, May 2016.

[7] M. Berioli, A. Molinaro, S. Morosi, and S. Scalise, "Aerospace communications for emergency applications," *Proc. IEEE*, vol. 99, no. 11, pp. 1922–1938, Nov. 2011.

[8] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

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

[10] J. Zhang, C.-K. Wen, S. Jin, X. Gao, and K.-K. Wong, "On capacity of large-scale MIMO multiple access channels with distributed sets of correlated antennas," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 133–148, Feb. 2013.

[11] A. Adhikary, J. Nam, J.-Y. Ahn, and G. Caire, "Joint spatial division and multiplexing—The large-scale array regime," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6441–6463, Oct. 2013.

[12] C. Sun, X. Q. Gao, S. Jin, M. Matthaiou, Z. Ding, and C. Xiao, "Beam division multiple access transmission for massive MIMO communications," *IEEE Trans. Commun.*, vol. 63, no. 6, pp. 2170–2184, Jun. 2015.

[13] Y.-H. Nam *et al.*, "Full-dimension MIMO (FD-MIMO) for next generation cellular technology," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 172–179, Jun. 2013.

[14] X. Li, S. Jin, H. A. Suraweera, J. Hou, and X. Gao, "Statistical 3-D beamforming for large-scale MIMO downlink systems over rician fading channels," *IEEE Trans. Commun.*, vol. 64, no. 4, pp. 1529–1543, Apr. 2016.

[15] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 160–171, Feb. 2013.

[16] Q. Zhang, S. Jin, K.-K. Wong, H. Zhu, and M. Matthaiou, "Power scaling of uplink massive MIMO systems with arbitrary-rank channel means," *IEEE J. Sel. Areas Commun.*, vol. 8, no. 5, pp. 966–981, Oct. 2014.

[17] H. Xie, F. Gao, S. Zhang, and S. Jin, "A unified transmission strategy for TDD/FDD massive MIMO systems with spatial basis expansion model," *IEEE Trans. Veh. Technol.*, vol. 66, no. 4, pp. 3170–3184, Apr. 2017.

[18] X. Xia, D. Zhang, K. Xu, W. Ma, and Y. Xu, "Hardware impairments aware transceiver for full-duplex massive MIMO relaying," *IEEE Trans. Signal Process.*, vol. 63, no. 24, pp. 6565–6580, Dec. 2015.

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

[20] A. Al-Hourani, S. Kandeepan, and A. Jamalipour, "Modeling air-to-ground path loss for low altitude platforms in urban environments," in *Proc. IEEE Global Commun. Conf.*, Austin, TX, USA, Dec. 2014, pp. 2898–2904.

[21] E. T. Michailidis, P. Theofilakos, and A. G. Kanatas, "Three-dimensional modeling and simulation of MIMO mobile-to-mobile via stratospheric relay fading channels," *IEEE Trans. Veh. Technol.*, vol. 62, no. 5, pp. 2014–2030, Jun. 2013.

[22] Z. Lian, L. Jiang, C. He, and Q. Xi, "A novel multiuser HAP-MIMO channel model based on birth-death process," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–5.

[23] Y. Hu, Y. Hong, and J. Evans, "Modelling interference in high altitude platforms with 3D LoS massive MIMO," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[24] A. Kammoun, H. Khanfir, Z. Altman, M. Debbah, and M. Kamoun, "Preliminary results on 3D channel modeling: From theory to standardization," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1219–1229, Jun. 2014.

[25] Q.-U.-A. Nadeem, A. Kammoun, M. Debbah, and M.-S. Alouini, "3D massive MIMO systems: Modeling and performance analysis," *IEEE Trans. Wireless Commun.*, vol. 14, no. 12, pp. 6926–6939, Dec. 2015.

[26] X. Li, S. Jin, X. Gao, and R. W. Heath, "Three-dimensional beamforming for large-scale FD-MIMO systems exploiting statistical channel state information," *IEEE Trans. Veh. Technol.*, vol. 65, no. 11, pp. 8992–9005, Nov. 2016.

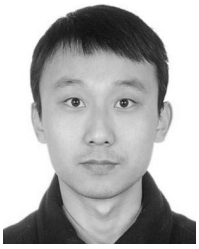
[27] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, Jr., "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014

[28] Y. Zeng and R. Zhang, "Millimeter wave MIMO with lens antenna array: A new path division multiplexing paradigm," *IEEE Trans. Commun.*, vol. 64, no. 4, pp. 1557–1571, Apr. 2016.

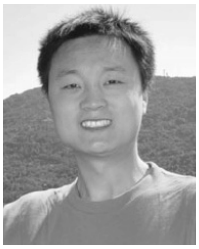
[29] L. You, X. Gao, X. G. Xia, N. Ma, and Y. Peng, "Pilot reuse for massive MIMO transmission over spatially correlated rayleigh fading channels," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 3352–3366, Jun. 2015.

[30] G. Caire, N. Jindal, M. Kobayashi, and N. Ravindran, "Multiuser MIMO achievable rates with downlink training and channel state feedback," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2845–2866, Jun. 2010.
[31] L. Liang, W. Xu, and X. Dong, "Low-complexity hybrid precoding in massive multiuser MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 3, no. 6, pp. 653–656, Dec. 2014.
[32] X. Xia, K. Xu, D. Zhang, Y. Xu, and Y. Wang, "Beam-domain full-duplex massive MIMO: Realizing co-time co-frequency uplink and downlink transmission in the cellular system," *IEEE Trans. Veh. Technol.*, to be published, doi: 10.1109/TVT.2017.2698160.
[33] Y. Wang, X. Xia, K. Xu, and A. Liu, "Location-assisted precoding for three-dimension massive sMIMO in air-to-ground transmission," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Atlanta, GA, USA, 2017.
[34] J. O. Smith, *Spectral Audio Signal Processing*, W3K Publishing, accessed on Dec. 2011. [Online]. Available: http://books.w3k.org/
[35] *Universal Mobile Telecommunications System (UMTS); Spatial Channel Model for Multiple Input Multiple Output (MIMO) Simulations, v.12.0.0*, 3GPP TR 25.996. Jun. 2012. [Online]. Available: http://www.3gpp.org
[36] C. Brunelli, H. Berg, and D. Guevorkian, "Approximating sine functions using variable-precision Taylor polynomials," in *Proc. IEEE Workshop Signal Process. Syst.*, Tampere, Finland, Oct. 2009, pp. 057–062.
[37] Y. Zhu, P. Y. Kam, and Y. Xin, "On the mutual information distribution of MIMO rician fading channels," *IEEE Trans. Commun.*, vol. 57, no. 5, pp. 1453–1462, May 2009.

YOUYUN XU (M'02–SM'11) was born in 1966. He received the Ph.D. degree in information and communication engineering from Shanghai Jiao Tong University (SJTU), China, in 1999. He is currently a Professor with the Nanjing Institute of Communication Engineering, China. He is also a part-time Professor with the Institute of Wireless Communication Technology, SJTU. He has more than 20 years professional experience teaching and researching in communication theory and engineering. His current research interests focus on new generation wireless mobile communication system (IMT-advanced and related), advanced channel coding and modulation techniques, multi-user information theory and radio resource management, wireless sensor networks, and cognitive radio networks. He is a Senior Member of the Chinese Institute of Electronics.



XIAOCHEN XIA was born in China, 1987. He received the B.S. degree in electronic science and technology from Tianjin University, China, in 2010, and the M.S. degree in communication and information systems from PLA University of Science and Technology (PLAUST) in 2013. He is currently pursuing the Ph.D. degree with the Institution of Communications Engineering, PLAUST. His research interests include relaying network, full-duplex communications, network coding, and MIMO techniques. He received the 2013 Master Degree Dissertation Award of Jiangsu Province, China.



KUI XU (M'13) was born in China. He received the B.S., M.S., and Ph.D. degrees from PLA University of Science and Technology (PLAUST), Nanjing, China, in 2004, 2006, and 2009, respectively. He is currently a Lecturer with the Wireless Communications Department, Institution of Communications Engineering, PLAUST. His research interests include multicarrier modulation, synchronization, signal processing in communications, network coding, and blind source separation.

YURONG WANG was born in China, 1990. She received the B.S. degree in computer science and technology from the Beijing Institute of Technology in 2013, and the M.S. degree in communication and information systems from PLA University of Science and Technology (PLAUST) in 2016. She is currently pursuing the Ph.D. degree with the Institution of Communications Engineering, PLAUST. Her research interests include full-duplex communication, massive MIMO systems, and broadband wireless communications.