Fully-Connected vs. Sub-Connected Hybrid Precoding Architectures for mmWave MU-MIMO

Xiaoshen Song, Thomas Kühne, and Giuseppe Caire
Communications and Information Theory Chair, Technische Universität Berlin, Germany

***Abstract*—Hybrid digital analog (HDA) beamforming has attracted considerable attention in practical implementation of millimeter wave (mmWave) multiuser multiple-input multiple-output (MU-MIMO) systems due to its low power consumption with respect to its digital baseband counterpart. The implementation cost, performance, and power efficiency of HDA beamforming depends on the level of connectivity and reconfigurability of the analog beamforming network. In this paper, we investigate the performance of two typical architectures for HDA MU-MIMO, i.e., the fully-connected (FC) architecture where each RF antenna port is connected to all antenna elements of the array; and the one-stream-per-subarray (OSPS) architecture where the RF antenna ports are connected to disjoint subarrays. We jointly consider the initial beam acquisition phase and data communication phase, such that the latter takes place by using the beam direction information obtained in the former phase. For each phase, we propose our own BA and precoding schemes that outperform the counterparts in the literature. We also evaluate the power efficiency of the two HDA architectures taking into account the practical hardware impairments, e.g., the power dissipation at different hardware components as well as the potential power backoff under typical power amplifier (PA) constraints. Numerical results show that the two architectures achieve similar sum spectral efficiency, but the OSPS architecture outperforms the FC case in terms of hardware complexity and power efficiency, only at the cost of a slightly longer time of initial beam acquisition.**

***Index Terms*—mmWave, hybrid, MIMO, sub-connected, fully-connected, beam alignment, spectral efficiency.**

I. Introduction

Millimeter wave (mmWave) multiuser multiple-input multiple-output (MU-MIMO) with large antenna arrays has been considered as a promising solution to meet the ever increasing data traffic for further 5G wireless communications [1]. Considering the unaffordable hardware cost of conventional full-digital baseband precoding, a combination of both digital and analog precoding, using a reduced number of RF chains, known as hybrid digital analog (HDA) structure, has been widely considered [1].

A large number of works have been dedicated to the optimization and performance characterization of HDA MU-MIMO architectures (e.g., see [1–4] and references therein). However, we observe that the current literature has some significant shortcomings. In particular 1) Many works consider only phase-shift control at the analog precoders [2–4].

X. Song is sponsored by the China Scholarship Council (201604910530). This work was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 779305 (SERENA).

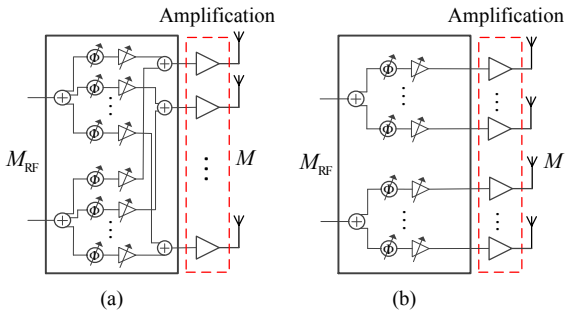

Fig. 1: Hybrid transmitter architectures with (a) fully-connected (FC), and (b) sub-connected with one-stream-per-subarray (OSPS).

This may somewhat reduce the hardware complexity, however, the signaling freedom is also drastically reduced. Particularly, it has been demonstrated in practical implementations that simultaneous amplitude and phase control with a low complexity and cost is feasible at mmWaves [5]. 2) Numerous works have ignored the hardware impairments [2] such as the power dissipation and non-linear distortion of the power amplifiers (PAs), which have an important effect on the signal processing and should not be neglected. 3) Lots of works investigate only the data communication phase and assume full channel state information (CSI) [2–4], i.e., they explicitly or implicitly assume that the precoder can be optimized using the channel coefficients seen at each antenna element, as if the signal from each antenna could be individually acquired. In contrast, however, due to the severe isotropic pathloss, mmWave communication requires an initial beam alignment (BA) phase to find the strongest narrow beam pair connecting each user equipment (UE) with the base station (BS). Since the number of RF chains in a HDA architecture (see Fig. 1) is much smaller than the number of antenna elements, it is impossible to obtain at once all the channel coefficients from one round of training signals, as commonly done with digital baseband schemes. Hence, only the effective channel along the beams acquired in the BA phase can be probed and measured.

In this paper, beyond the above fundamental limitations in the literature, we evaluate the performance of two typical transmitter architectures. On one hand, a fully-connected (FC) architecture where each RF chain (antenna port) is connected to all antenna elements of the array (Fig. 1 (a)). At the other extreme, a one-stream-per-subarray (OSPS) architecture where the RF chains are connected to disjoint subarrays (Fig. 1 (b)). We jointly consider the initial BA phase, data communication, and practical hardware impairments. In particular, we propose

our own BA/precoding scheme and evaluation model in order to provide a useful analysis and comparison framework for mmWave system design.

II. Channel Model

We consider a system formed by a BS equipped with a uniform linear array (ULA) with M antennas and M_{RF} RF chains, serving simultaneously $K = M_{\text{RF}}$ UEs, each of which has also a ULA with N antennas and N_{RF} RF chains. The propagation channel between the BS and the k-th UE, $k \in [K]$, consists of $L_k \ll \max\{M, N\}$ multi-path components. Accordingly, the baseband equivalent impulse response of the channel at time slot s reads

$$\begin{aligned}\mathsf{H}_{k,s}(t,\tau) &= \sum_{l=1}^{L_k} \rho_{k,s,l} e^{j2\pi\nu_{k,l}t} \mathbf{a}_{\text{R}}(\phi_{k,l})\mathbf{a}_{\text{T}}(\theta_{k,l})^{\mathsf{H}} \delta(\tau - \tau_{k,l}) \\ &= \sum_{l=1}^{L_k} \mathsf{H}_{k,s,l}(t)\delta(\tau - \tau_{k,l}), \end{aligned} \tag{1}$$

where $\mathsf{H}_{k,s,l}(t) := \rho_{k,s,l} e^{j2\pi\nu_{k,l}t} \mathbf{a}_{\text{R}}(\phi_{k,l})\mathbf{a}_{\text{T}}(\theta_{k,l})^{\mathsf{H}}$, $(\phi_{k,l}, \theta_{k,l}, \tau_{k,l}, \nu_{k,l})$ denote the angle of arrival (AoA), angle of departure (AoD), delay, and Doppler shift of the l-th component, and $\delta(\cdot)$ denotes the Dirac delta function. The vectors $\mathbf{a}_{\text{T}}(\theta_{k,l}) \in \mathbb{C}^D$ and $\mathbf{a}_{\text{R}}(\phi_{k,l}) \in \mathbb{C}^N$ are the array response vectors of the BS and UE at AoD $\theta_{k,l}$ and AoA $\phi_{k,l}$, respectively, with elements given by

$$[\mathbf{a}_{\text{T}}(\theta)]_d = e^{j(d-1)\pi\sin(\theta)},\ d \in [D], \tag{2a}$$

$$[\mathbf{a}_{\text{R}}(\phi)]_n = e^{j(n-1)\pi\sin(\phi)},\ n \in [N], \tag{2b}$$

where $D = M$ for Fig. 1(a) and $D = \frac{M}{M_{\text{RF}}}$ for Fig. 1(b). Here we assume that the spacing of the ULA antennas equals to the half of the wavelength. We adopt a block fading model, i.e., the channel gains $\rho_{k,s,l}$ remain invariant over the channel *coherence time* Δt_c but change i.i.d. randomly across different Δt_c. Since each scatterer in practice is a superposition of many smaller components that have (roughly) the same AoA-AoD and delay, we assume a general Rice fading model given by

$$\rho_{k,s,l} \sim \sqrt{\gamma_{k,l}} \left(\sqrt{\frac{\eta_{k,l}}{1+\eta_{k,l}}} + \frac{1}{\sqrt{1+\eta_{k,l}}} \check{\rho}_{k,s,l} \right), \tag{3}$$

where $\gamma_{k,l}$ denotes the overall multi-path component strength, $\eta_{k,l} \in [0, \infty)$ indicates the strength ratio between the line-of-sight (LOS) and the non-LOS (NLOS) components, and $\check{\rho}_{k,s,l} \sim \mathcal{CN}(0,1)$ is a zero-mean unit-variance complex Gaussian random variable. In particular, $\eta_{k,l} \to \infty$ indicates a pure LOS path while $\eta_{k,l} = 0$ indicates a pure NLOS path, affected by standard Rayleigh fading.

Following the *beamspace representation* as in [6], we obtain an approximate finite-dimensional representation of the channel response (1) with respect to the discrete dictionary in the AoA-AoD (beam) domain defined by the quantized angles

$$\Phi := \{\check{\phi} : (1 + \sin(\check{\phi}))/2 = \frac{n-1}{N},\ n \in [N]\}, \tag{4a}$$

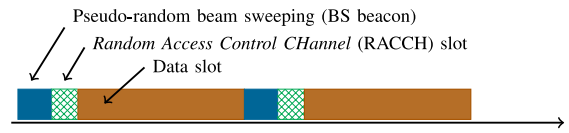

Fig. 2: Illustration of the frame structure in the underlying system.

$$\Theta := \{\check{\theta} : (1 + \sin(\check{\theta}))/2 = \frac{m-1}{M}, m \in [M]\}, \tag{4b}$$

with corresponding array response vectors $\mathcal{A}_{\text{R}} := \{\mathbf{a}_{\text{R}}(\check{\phi}) : \check{\phi} \in \Phi\}$ and $\mathcal{A}_{\text{T}} := \{\mathbf{a}_{\text{T}}(\check{\theta}) : \check{\theta} \in \Theta\}$. For ULAs as considered in this paper, the dictionaries $\mathcal{A}_{\text{R}}$ and $\mathcal{A}_{\text{T}}$, after suitable normalization, yield to the discrete Fourier transform (DFT) matrices $\mathbf{F}_N \in \mathbb{C}^{N\times N}$ and $\mathbf{F}_D \in \mathbb{C}^{D\times M}$ with elements

$$[\mathbf{F}_N]_{n,n'} = \frac{1}{\sqrt{N}} e^{j2\pi(n-1)(\frac{n'-1}{N} - \frac{1}{2})}, n, n' \in [N], \tag{5a}$$

$$[\mathbf{F}_D]_{d,d'} = \frac{1}{\sqrt{M}} e^{j2\pi(d-1)(\frac{d'-1}{M} - \frac{1}{2})}, d \in [D], d' \in [M]. \tag{5b}$$

Consequently, the beam-domain channel representation reads

$$\check{\mathsf{H}}_{k,s}(t,\tau) = \mathbf{F}_N^{\mathsf{H}} \mathsf{H}_{k,s}(t,\tau) \mathbf{F}_D = \sum_{l=1}^{L_k} \check{\mathsf{H}}_{k,s,l}(t)\delta(\tau - \tau_{k,l}), \tag{6}$$

where $\check{\mathsf{H}}_{k,s,l}(t) := \mathbf{F}_N^{\mathsf{H}} \mathsf{H}_{k,s,l}(t) \mathbf{F}_D$. It is well-known (e.g., see [7] and references therein) that, as M and N increase, the DFT basis provides a very sparse channel representation.

III. Beam Acquisition and Data Transmission

Fig. 2 illustrates the considered frame structure which consists of three parts [6]: the beacon slot, the random access control channel (RACCH) slot, and the data slot. As illustrated in our previous work [6], in the initial acquisition phase the measurements are collected at the UEs from downlink beacon slots broadcasted by the BS. Each UE selects its strongest AoA as the beamforming direction for possible data transmission. During the RACCH slot, the BS stays in listening mode such that each UE sends a beamformed packet to the BS. This packet contains basic information such as the UE ID and the beam indices of the selected AoDs. The BS responds with an acknowledgment data packet in the data subslot of a next frame. From this moment on the BS and the UE are connected in the sense that, if the procedure is successful, they have achieved BA. In other words, they can communicate by aligning their beams along a mutlipath component with AoA-AoD $(\phi_{k,l}, \theta_{k,l})$ and strong coefficient $\rho_{k,l}$.

The details of the BA algorithm and its performance are given in [7] for a frequency-domain based variant of the problem, and in [6] for a time-domain single-carrier variant of the problem. In both cases, we have shown that the proposed scheme aligns the beams along the strongest multipath component with probability that tends to 1 as the number of beacon slots (measurements) increases, and that the probability of collisions or errors on the RACCH protocol information exchange is negligible when the alignment directions have been correctly found.

IV. System Analysis

A. Hardware Impairments

We assume that each analog path has simultaneous amplitude and phase control as shown in Fig. 1. Let $\tilde{\mathbf{x}} = [\tilde{x}_1, \cdots, \tilde{x}_M] \in \mathbb{C}^M$ denote the beamformed signal[1] given by

$$\tilde{\mathbf{x}} = \sqrt{\alpha_{\text{com}}}\tilde{\mathbf{U}} \cdot \sqrt{\alpha_{\text{div}}}\mathbf{x}, \tag{7}$$

where $\mathbf{x} = [x_1, \cdots, x_{M_{\text{RF}}}] \in \mathbb{C}^{M_{\text{RF}}}$ is the transmit complex symbol vector, with $\mathbb{E}[|x_k|^2] = \epsilon$, $k \in [M_{\text{RF}}]$. α_{div} results from to the signal splitting with $\alpha_{\text{div}} = \frac{1}{M}$ for Fig. 1 (a) and $\alpha_{\text{div}} = \frac{M_{\text{RF}}}{M}$ for Fig. 1 (b). α_{com} models the power dissipation factor of the combiners corresponding to their S-parameters as in [3] with $\alpha_{\text{com}} = \frac{1}{M_{\text{RF}}}$ for Fig. 1 (a) and $\alpha_{\text{com}} = 1$ for Fig. 1 (b). $\tilde{\mathbf{U}} \in \mathbb{C}^{M \times M_{\text{RF}}}$ denotes the overall beamforming coefficients, given by

$$[\tilde{\mathbf{u}}_1, \tilde{\mathbf{u}}_2, \cdots, \tilde{\mathbf{u}}_{M_{\text{RF}}}] \quad \text{and} \quad \begin{bmatrix} \tilde{\mathbf{u}}_1 & 0 & \dots & 0 \\ 0 & \tilde{\mathbf{u}}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \tilde{\mathbf{u}}_{M_{\text{RF}}} \end{bmatrix} \tag{8}$$

for Fig. 1 (a) and Fig. 1 (b), respectively. To meet the total input constraint, in Fig. 1 (a) we have $\tilde{\mathbf{u}}_k \in \mathbb{C}^M$ with $\|\tilde{\mathbf{u}}_k\|^2 = M$, whereas in Fig. 1 (b) we assume $\tilde{\mathbf{u}}_k \in \mathbb{C}^{\frac{M}{M_{\text{RF}}}}$ with $\|\tilde{\mathbf{u}}_k\|^2 = \frac{M}{M_{\text{RF}}}$, $k \in [M_{\text{RF}}]$. Based on (7), the sum-power of the beamformed signal $\tilde{\mathbf{x}}$ can be written as

$$\tilde{P} = \mathbb{E}[\tilde{\mathbf{x}}^{\mathsf{H}}\tilde{\mathbf{x}}] = \alpha_{\text{com}}\alpha_{\text{div}} \cdot \text{tr}\left(\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathsf{H}}\tilde{\mathbf{U}}^{\mathsf{H}}\tilde{\mathbf{U}}]\right). \tag{9}$$

Consequently, the sum-power for the FC architecture of Fig. 1 (a) and for the OSPS architecture of Fig. 1 (b) reads $\tilde{P}_{\text{FC}} = \epsilon M_{\text{RF}}\frac{1}{M_{\text{RF}}}$ and $\tilde{P}_{\text{OSPS}} = \epsilon M_{\text{RF}}$, respectively. In order to compensate the additional combiner power dissipation in Fig. 1 (a), the transmitter should either boost the input signal as $M_{\text{RF}}\mathbf{x}$ or choose PAs with larger gain for the amplification stage. We consider the former approach and include $(\alpha_{\text{com}}, \alpha_{\text{div}})$ into the column-wise normalized beamforming matrix $\widetilde{\mathbf{U}}$, such that we can write (7) as

$$\tilde{\mathbf{x}} = \widetilde{\mathbf{U}}\mathbf{x}. \tag{10}$$

The beamformed signal then goes through the amplification stage, where at each antenna branch a PA amplifies the signal before transmission. We assume that the PAs in different antenna branches have the same input-output relation. For any given antenna in the transmitter array, let P_{rad} denote the radiated power of the antenna, and P_{cons} denote the consumed power by the corresponding PA including both the radiated power and the dissipated power. Following the approach in [8], the power consumed by the PA reads

$$P_{\text{cons}} = \frac{\sqrt{P_{\text{max}}}}{\eta_{\text{max}}}\sqrt{P_{\text{rad}}}, \tag{11}$$

where P_{max} is the maximum output power of the PA with $P_{\text{rad}} \leq P_{\text{max}}$, and η_{max} is the maximum efficiency of the PA. Considering that the PAs are often the predominant power consumption part, we define η_{eff} given by

$$\eta_{\text{eff}} = \frac{P_{\text{rad}}}{P_{\text{cons}}} \tag{12}$$

as the metric to effectively compare the power efficiency of the two transmitter architectures in Fig. 1.

Due to the superposition of multiple beamforming vectors (particularly in the FC case) and the potentially high peak to average power ratio (PAPR) of the time-domain transmit waveform x_k (particularly with orthogonal frequency division multiplexing), the input power for some individual PA may exceed its saturation limit. This would result in non-linear distortion and even the disruption of the whole transmission. To compare the two transmitter architectures and ensure that all the underlying M PAs simultaneously work in their linear range, we generally have two options:

Option I: All transmitters utilize the same PA but apply a different input back-off $\alpha_{\text{off}} \in (0, 1]$, such that the peak power of the radiated signal is smaller than P_{max}. As a reference, we denote by $(P_{\text{rad},0}, \eta_{\text{max},0})$ as the parameters of a reference PA under the reference precoding/beamforming strategy with a power backoff factor $\alpha_{\text{off},0}$ (as illustrated later in Section V). For different scenarios (with certain α_{off}) the effective radiated power and the consumed power read $P_{\text{rad}} = \frac{\alpha_{\text{off}}}{\alpha_{\text{off},0}}P_{\text{rad},0}$, $P_{\text{cons}} = \frac{\sqrt{P_{\text{max},0}}}{\eta_{\text{max},0}}\sqrt{P_{\text{rad}}}$. The transmitter efficiency is given by

$$\eta_{\text{eff}} = \frac{P_{\text{rad}}}{P_{\text{cons}}} = \frac{\sqrt{P_{\text{rad}}} \cdot \eta_{\text{max},0}}{\sqrt{P_{\text{max},0}}}. \tag{13}$$

Option II: We choose to deploy different PAs for different transmitter architectures, with a maximum output power given by $P_{\text{max}} = \frac{\alpha_{\text{off},0}}{\alpha_{\text{off}}}P_{\text{max},0}$, where α_{off} has the same value as in *Option I*. Consequently, the effective radiated power and the consumed power of the underlying PA read $P_{\text{rad}} = P_{\text{rad},0}$, $P_{\text{cons}} = \frac{\sqrt{P_{\text{max},0} \cdot \alpha_{\text{off},0}/\alpha_{\text{off}}}}{\eta_{\text{max}}}\sqrt{P_{\text{rad}}}$. The transmitter efficiency is given by

$$\eta_{\text{eff}} = \frac{P_{\text{rad}}}{P_{\text{cons}}} = \frac{\sqrt{P_{\text{rad}}} \cdot \eta_{\text{max}}}{\sqrt{P_{\text{max},0} \cdot \alpha_{\text{off},0}}} \cdot \sqrt{\alpha_{\text{off}}}. \tag{14}$$

Note that the characteristics (P_{max} and η_{max}) of different PAs highly depend on the operation frequency, implementation, and technology. Aiming at illustrating how to apply the proposed analysis framework in practical system design, we will exemplify a set of PA parameters in Section V to evaluate the efficiency η_{eff} of the two architectures in Fig. 1. However, in the following derivations for the BA and data communication, otherwise stated, we will assume a single-carrier (SC) modulation and a fixed total radiated power constraint denoted by $P_{\texttt{tot}}$, where all the underlying PAs work in their linear range with an identical scalar gain.

B. Initial BA Phase

As discussed in Section I, communication at mmWaves requires narrow beams via large antenna array beamforming to overcome the severe signal attenuation. In this section,

[1] Here for notation simplicity, we neglected the signal time-domin index t.

we provide a brief description of our recently proposed time-domain BA scheme and refer to [6] for more details.

In short, we assume that the BS broadcasts its pilot signals periodically over the beacon slots according to a pseudo-random beamforming codebook, which is known to all the UEs in the system. We assign a unique Pseudo Noise (PN) sequence as the pilot signal to each RF chain at the BS such that different pilot streams are separable at the UE. Meanwhile, each UE independently collects its measurements to estimate its strong AoA-AoD combinations. Taking the k-th UE for example, let $\mathbf{\Gamma}_k$ denote the $N \times M$ matrix with elements corresponding to the beam-domain second-order statistics of the channel coefficients, given by

$$[\mathbf{\Gamma}_k]_{n,m} \propto \sum_{l=1}^{L_k} \mathbb{E}\left[\left|[\check{\mathsf{H}}_{k,s,l}(t,\tau_{k,l})]_{n,m}\right|^2\right]. \tag{15}$$

As confirmed by many measurement campaigns that the propagation at mmWaves is dominated by the LoS path (when it exists) and a few strong multipath components [5]. Hence, the matrix $\mathbf{\Gamma}_k$ is generally all "almost zero", but for a few large positive elements. Accordingly, over T beacon slots the UE obtains a total number of $M_{\text{RF}}N_{\text{RF}}T$ equations, which can be written in the form

$$\mathbf{q}_k = \mathbf{B}_k \cdot \text{vec}(\mathbf{\Gamma}_k) + \zeta(P_{\texttt{tot}}) \cdot \mathbf{1} + \mathbf{w}_k, \tag{16}$$

where $\mathbf{q}_k \in \mathbb{R}^{M_{\text{RF}}N_{\text{RF}}T}$ consists of all the $M_{\text{RF}}N_{\text{RF}}T$ statistical power measurements, $\mathbf{B}_k \in \mathbb{R}^{M_{\text{RF}}N_{\text{RF}}T \times MN}$ is uniquely defined by the pseudo-random beamforming codebook of the BS and the local beamforming codebook of the k-th UE, $\zeta(P_{\texttt{tot}})$ denotes a constant whose value is a function of the total radiated power, and $\mathbf{w}_k \in \mathbb{R}^{M_{\text{RF}}N_{\text{RF}}T}$ denotes the residual measurement fluctuations. As discussed in [6], with the non-negative constraint of $\mathbf{\Gamma}_k$, a simple Least Squares (LS)

$$\mathbf{\Gamma}_k^\star = \underset{\mathbf{\Gamma}_k \in \mathbb{R}_+^{N\times M}}{\arg\min} \|\mathbf{B} \cdot \text{vec}(\mathbf{\Gamma}_k) + \zeta(P_{\texttt{tot}}) \cdot \mathbf{1} - \mathbf{q}_k\|^2 \tag{17}$$

is sufficient to impose the sparsity of the solution $\mathbf{\Gamma}_k^\star$. We assume a success of the BA if the largest component in $\mathbf{\Gamma}_k^\star$ coincides with the actual strongest path of the k-th UE. More details can be found in [6].

C. Data Communication Phase

The number of RF chains M_{RF} at the BS determines the maximum number of downlink data streams that the BS can send. We assume that the BS simultaneously schedules $K = M_{\text{RF}}$ UEs, which are selected by a simple directional scheduler [9]. Namely, the selected K UEs have similar power profiles and their strongest AoDs in the downlink are well separated. Denoted by $\mathbf{u}_k$ as the normalized transmit beamforming vector for the k-th UE at the BS and $\mathbf{v}_k$ as the normalized receive beamforming vector at the k-th UE, with an effective radiated power $P_k = \frac{P_{\texttt{tot}}}{M_{\text{RF}}}$ to maintain the total radiated power constraint, the received signal at the k-th UE can be written as

$$\begin{aligned} y_k(t) =& \mathbf{v}_k^{\mathsf{H}} \sum_{k'=1}^{K} \sqrt{P_{k'}} \mathsf{H}_{k,s}(t,\tau) \circledast (\mathbf{u}_{k'} x_{k'}(t)) + z_k(t) \\ =& \sqrt{P_k}\left(\mathbf{v}_k^{\mathsf{H}} \mathsf{H}_{k,s}(t,\tau)\mathbf{u}_k\right) \circledast x_k(t) + z_k(t) \\ &+ \sum_{k'\neq k} \sqrt{P_{k'}}\left(\mathbf{v}_k^{\mathsf{H}} \mathsf{H}_{k,s}(t,\tau)\mathbf{u}_{k'}\right) \circledast x_{k'}(t) \end{aligned} \tag{18}$$

where $f(t) \circledast g(t) = \int f(\tau) g(t-\tau) d\tau$ denotes the convolution operation. As we can see, the first term in (18) corresponds to the desired signal at the k-th UE, whereas the last two terms correspond to the noise and interference, respectively. By substituting (1) into (18), the received signal reads

$$\begin{aligned} y_k(t) =& \sum_{l=1}^{L_k} \sqrt{P_k} \mathbf{v}_k^{\mathsf{H}} \mathsf{H}_{k,s,l}(t) \mathbf{u}_k x_k(t-\tau_{k,l}) + z_k(t) \\ &+ \sum_{k'\neq k} \sum_{l=1}^{L_{k'}} \sqrt{P_{k'}} \mathbf{v}_k^{\mathsf{H}} \mathsf{H}_{k',s,l}(t) \mathbf{u}_{k'} x_{k'}(t-\tau_{k',l}), \end{aligned} \tag{19}$$

where $x_k(t)$ denotes the unit-power transmit signal, $z_k(t) \sim \mathcal{CN}(0, N_0 B)$ denotes the continuous-time complex additive white Gaussian noise (AWGN) with a power spectral density (PSD) of N_0 Watt/Hz, and B denotes the effective bandwidth. By treating the multi-user interference as noise at each UE the asymptotic spectral efficiency of the k-th UE is given by

$$R_k = \mathbb{E}\left[\log_2\left(1 + \frac{P_k |\sum_{l=1}^{L_k} \mathbf{v}_k^{\mathsf{H}} \mathsf{H}_{k,s,l}(t) \mathbf{u}_k|^2}{|\sum\sum_{k'\neq k}^{L_{k'}}_{l=1} \sqrt{P_{k'}} \mathbf{v}_k^{\mathsf{H}} \mathsf{H}_{k',s,l}(t) \mathbf{u}_{k'}|^2 + |z_k(t)|^2}\right)\right], \tag{20}$$

and the sum spectral efficiency reads $R_{\text{sum}} = \sum_{k=1}^{K} R_k$.

We claim that the beamforming vectors corresponding to each UE in the data communication phase are based on the outcome of the BA in Section IV-B. More precisely, assume that after a BA procedure the strongest component in $\mathbf{\Gamma}_k^\star$ corresponds to the l_k-th multi-path component in $\check{\mathsf{H}}_{k,s}(t,\tau)$ between the BS and the k-th UE. To simplify the practical implementation, we assume that the k-th UE decodes its data along the estimated strongest direction, given by

$$\mathbf{v}_k = \mathbf{F}_N \check{\mathbf{v}}_k, \tag{21}$$

where $\check{\mathbf{v}}_k \in \mathbb{C}^N$ is an all-zero vector with a 1 at the component corresponding to the AoA of the l_k-th scatterer. At the BS we assume that the BS communicates with the k-th UE along its top-p beams with respect to the AoA given by $\mathbf{v}_k$. With $p \geq 1$, the BS can choose wider beams to handle the potential mobility or blockage. Define $\mathbf{U}_k \in \mathbb{C}^{D\times p}$, each column of which corresponds to one of the p AoD beamforming directions, given by

$$\begin{aligned} \mathbf{U}_k &= [\mathbf{u}_{k,1}, \mathbf{u}_{k,2}, ..., \mathbf{u}_{k,p}] \\ &= \mathbf{F}_D \cdot [\check{\mathbf{u}}_{k,1}, \check{\mathbf{u}}_{k,2}, ..., \check{\mathbf{u}}_{k,p}], \end{aligned} \tag{22}$$

where $\check{\mathbf{u}}_{k,i} \in \mathbb{C}^M$, $i \in [p]$, is an all-zero vector with a 1 at the component corresponding to the i-th strongest AoD direction of the k-th UE, and where $D = M$ for the FC architecture, otherwise $D = \frac{M}{M_{\text{RF}}}$ for the OSPS case.

To formulate the hybrid precoding problem, we re-write everything in a matrix-multiplication format. Let $\mathbf{x}(t) =$

$\text{diag}(\sqrt{P_1}, \sqrt{P_2}, ..., \sqrt{P_K}) \cdot [x_1(t), x_2(t), ..., x_K(t)]^{\mathsf{T}} \in \mathbb{C}^K$ denote the transmit signal vector and $\overline{\mathsf{H}}_s(t,\tau)$ denote the aggregated channel for all the K UEs given by

$$\overline{\mathsf{H}}_s(t,\tau) = \left[\mathsf{H}_{1,s}(t,\tau)^{\mathsf{T}}, \mathsf{H}_{2,s}(t,\tau)^{\mathsf{T}}, \cdots, \mathsf{H}_{K,s}(t,\tau)^{\mathsf{T}}\right]^{\mathsf{T}}, \quad (23)$$

where $\mathsf{H}_{k,s}(t,\tau)$, $k \in [K]$, is given in (1). We define $\mathbf{V} \in \mathbb{C}^{NK \times K}$ as the receive beamforming matrix given by

$$\begin{aligned} \mathbf{V} &= \text{diag}(\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_K) \\ &= (\mathbf{I}_K \otimes \mathbf{F}_N) \cdot \text{diag}(\check{\mathbf{v}}_1, \check{\mathbf{v}}_2, ..., \check{\mathbf{v}}_K), \end{aligned} \quad (24)$$

where $\mathbf{I}_K$ denotes the $K \times K$ identity matrix, and $\otimes$ represents the Kronecker product. Let $\overline{\mathbf{U}} \in \mathbb{C}^{D \times pK}$ denote the analog precoding vector support given by

$$\begin{aligned} \overline{\mathbf{U}} &= [\mathbf{U}_1, \mathbf{U}_2, ..., \mathbf{U}_K] \\ &= \mathbf{F}_D \cdot [\check{\mathbf{u}}_{1,1}, ..., \check{\mathbf{u}}_{1,p}, ..., \check{\mathbf{u}}_{K,1} ..., \check{\mathbf{u}}_{K,p}], \end{aligned} \quad (25)$$

and $\mathbf{A}_B = [\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_K] \in \mathbb{C}^{pK \times K}$ denote the baseband precoding matrix.[2] The overall precoding matrix $\mathbf{U} \in \mathbb{C}^{D \times K}$ at the BS can be written as

$$\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_K] = \overline{\mathbf{U}} \cdot \mathbf{A}_B. \quad (26)$$

To meet the total radiated power constraint, the coefficients in (26) are normalized as $\|\mathbf{u}_k\| = \|\overline{\mathbf{U}} \cdot \mathbf{a}_k\| = 1$. As a result, the receive signal $\mathbf{y}(t) = [y_1(t), y_2(t), ..., y_K(t)]^{\mathsf{T}} \in \mathbb{C}^K$ reads

$$\begin{aligned} \mathbf{y}(t) &= \mathbf{V}^{\mathsf{H}} \cdot \overline{\mathsf{H}}_s(t,\tau) \circledast (\mathbf{U} \cdot \mathbf{x}(t)) + \mathbf{z}(t) \\ &= \left(\widetilde{\mathsf{H}}_s(t,\tau) \cdot \mathbf{A}_B\right) \circledast \mathbf{x}(t) + \mathbf{z}(t), \end{aligned} \quad (27)$$

where $\mathbf{z}(t) \in \mathbb{C}^K$ denotes the noise, and

$$\widetilde{\mathsf{H}}_s(t,\tau) = \mathbf{V}^{\mathsf{H}} \cdot \overline{\mathsf{H}}_s(t,\tau) \cdot \overline{\mathbf{U}} \quad (28)$$

represents the $K \times (p\,K)$-lower-dimensional effective channel. The effective channel can be easily estimated over $p \cdot K$ additional sub-slots in the uplink, on the condition that a successful BA procedure in Section IV-B has been achieved[3].

1) Beam Steering (BST) Scheme

In the beam steering (BST) scheme, we assume that the BS simply steers K data streams towards the K strongest AoDs, i.e., we have $p = 1$ in (22). As illustrated in Section II, the underlying beam indices are estimated and fed back from the corresponding K UEs. More preciously, assume that after a BA procedure as in Section IV-B, the strongest component (top first) in $\mathbf{\Gamma}_k^\star$ (17) corresponds to the l_k-th multi-path component in $\check{\mathsf{H}}_{k,s}(t,\tau)$. Consequently, the beamforming vector for the k-th UE at the BS is given by

$$\mathbf{u}_{k,1} = \mathbf{F}_D \cdot \check{\mathbf{u}}_{k,1}, \quad (29)$$

Further, the analog precoding vector support $\overline{\mathbf{U}}^{\text{BST}}$(25) and the baseband precoding matrix $\mathbf{A}_B^{\text{BST}}$ are given by $\overline{\mathbf{U}}^{\text{BST}} = [\mathbf{u}_{1,1}, \mathbf{u}_{2,1}, ..., \mathbf{u}_{K,1}]$ and $\mathbf{A}_B^{\text{BST}} = \mathbf{I}_K$, respectively. In this case, an additional uplink channel estimation of $\widetilde{\mathsf{H}}_s(t,\tau)$ can be omitted. The eventual $D \times K$ BST precoder in (26) reads

$$\mathbf{U}^{\text{BST}} = \overline{\mathbf{U}}^{\text{BST}} \cdot \mathbf{A}_B^{\text{BST}} = \overline{\mathbf{U}}^{\text{BST}}. \quad (30)$$

2) Baseband Zeroforcing (BZF) Scheme

In this scheme, we consider ZF precoding for potential multi-user interference cancellation. More precisely, we assume that after a BA phase as in Section IV-B, each UE steers its beam towards the estimated strongest AoA $\mathbf{v}_k$ (21). Meanwhile, each UE feeds back to the BS the AoD information of its top-p strongest paths along the direction $\mathbf{v}_k = \mathbf{F}_N \check{\mathbf{v}}_k$, where $1 \leq p \ll M$. Let $\mathbf{r}_k \in \mathbb{R}_+^M$ denote the non-negative second-order channel statistics corresponding to the k-th UE, given by

$$\mathbf{r}_k = (\mathbf{\Gamma}_k^\star)^{\mathsf{H}} \cdot \check{\mathbf{v}}_k, \quad (31)$$

where $\mathbf{\Gamma}_k^\star$ is given in (17). Let the elements in $\mathbf{r}_k$ be arranged in non-increasing order in terms of their strengths, i.e., $\mathbf{r}_k[m_1^k] \geq \mathbf{r}_k[m_2^k] \geq ...\mathbf{r}_k[m_p^k] \geq ... \geq \mathbf{r}_k[m_M^k]$, and define $\mathcal{A}_k = \{m_1^k, m_2^k, ..., m_p^k\}$ as the beam index set of the top-p strongest elements in $\mathbf{r}_k$. We assume that each UE feeds back its beam index set $\mathcal{A}_k$ to the BS through the RACCH slots as illustrated in Fig. 2. Consequently, the analog precoding vector support $\overline{\mathbf{U}}^{\text{ZF}}$ at the BS can be written as

$$\overline{\mathbf{U}}^{\text{ZF}} = [\mathbf{f}_{D,m_1^1}, ..., \mathbf{f}_{D,m_p^1}, ..., \mathbf{f}_{D,m_1^K}, ..., \mathbf{f}_{D,m_p^K}], \quad (32)$$

where $\mathbf{f}_{D,i}$ denotes the i-th column of the DFT matrix $\mathbf{F}_D$. Substituting (32) into (28), the effective channel $\widetilde{\mathsf{H}}_s(t,\tau)$ reads

$$\widetilde{\mathsf{H}}_s(t,\tau) = \mathbf{V}^{\mathsf{H}} \cdot \overline{\mathsf{H}}_s(t,\tau) \cdot \overline{\mathbf{U}}^{\text{ZF}}, \quad (33)$$

which can be estimated through an "exhaustive" procedure with orthogonal uplink pilots at the cost of $(p \cdot K \ll MN)$ sub-slots. As a result, the baseband precoding matrix $\mathbf{A}_B^{\text{ZF}}$ can be written as

$$\mathbf{A}_B^{\text{ZF}} = \widetilde{\mathsf{H}}_s(t,\tau)^{\mathsf{H}} \cdot \left(\widetilde{\mathsf{H}}_s(t,\tau)\widetilde{\mathsf{H}}_s(t,\tau)^{\mathsf{H}}\right)^{-1} \cdot \Delta^{\text{ZF}}, \quad (34)$$

where $\Delta^{\text{ZF}} \in \mathbb{R}_+^{K \times K}$ is a diagonal matrix, taking into account the total radiated power constraint. The eventual BZF precoder is then given by

$$\mathbf{U}^{\text{ZF}} = \overline{\mathbf{U}}^{\text{ZF}} \cdot \mathbf{A}_B^{\text{ZF}}. \quad (35)$$

In the following section, we will compare the asymptotic sum spectral efficiency in terms of different transmitter architectures. To effectively capture the channel quality before BA, we also define the SNR before beamforming (BBF) by

$$\mathsf{SNR}_{\text{BBF}} = \frac{P_{\mathtt{tot}} \sum_{l=1}^{L} \gamma_l}{N_0 B}. \quad (36)$$

This is the SNR obtained when a single pilot stream ($M_{\text{RF}} = 1$) is transmitted through a single BS antenna and is received at a single UE antenna (isotropic transmission) via a single RF chain ($N_{\text{RF}} = 1$) over the whole bandwidth B.

[2]We call $\mathbf{A}_B$ as the baseband precoding matrix just for the expression consistency with most existing hybrid papers, which consider only-phase-control at the analog part (see Section I). In this paper, we consider both phase and amplitude control for each analog path, hence, $\mathbf{A}_B$ is actually applied at the analog part, while the real baseband precoding is just an identity matrix, which is ignored in this paper for notation simplicity.

[3]We use channel reciprocity and standard uplink orthogonal pilot transmission for the lower-dimensional effective channel estimation of $\widetilde{\mathsf{H}}_s(t,\tau)$.

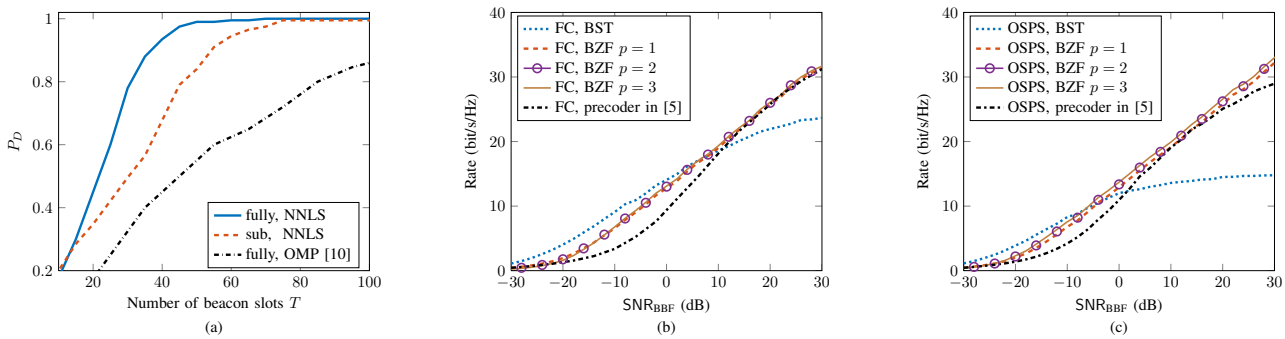

Fig. 3: (a) Detection probability P_D of different transmitter architectures vs. the training overhead, for the initial BA phase with $\mathsf{SNR}_{\mathrm{BBF}} = -20$ dB. (b) The sum spectral efficiency of the FC architecture vs. increasing $\mathsf{SNR}_{\mathrm{BBF}}$, for the data communication phase with different precoding schemes. (c) The sum spectral efficiency of the OSPS architecture vs. increasing $\mathsf{SNR}_{\mathrm{BBF}}$, for the data communication phase with different precoding schemes.

V. Numerical Results

We consider a system with a BS using $M = 32$ antennas and $M_{\mathrm{RF}} = 2$ RF chains and each UE using $N = 16$ antennas and $N_{\mathrm{RF}} = 1$ RF chain. The system is assumed to work at $f_0 = 40$ GHz with a maximum available bandwidth of $B = 0.8$ GHz. We assume the channel for each UE contains $L_k = 3$ paths given by $(\gamma_{k,1} = 1, \eta_{k,1} = 100)$, $(\gamma_{k,2} = 0.6, \eta_{k,2} = 10)$, and $(\gamma_{k,3} = 0.6, \eta_{k,3} = 0)$ as defined in (3). The strongest paths of the simultaneously scheduled UEs are well separated in the beam domain, while all the less strong paths of each UE are randomly distributed. In the following, we will compare the performance of the two transmitter architectures in Fig. 1.

A. Training Efficiency for the Initial BA Phase

Let P_D denote the detection probability, i.e., the probability of finding the strongest AoA-AoD pair between the BS and a generic UE. As illustrated in Fig. 3 (a), due to the fact that the OSPS architecture has lower angular resolution and encounters larger sidelobe power leakage than the FC case, the former requires moderately ~ 20 more beacon slots than the latter for $P_D \geq 0.95$. We also simulate a recent time-domain BA algorithm based on [10] which focuses on estimating the instantaneous channel coefficients with an orthogonal matching pursuit (OMP) technique. As we can see, for both transmitter architectures the proposed BA scheme requires much less training overhead than that in [10], implying its advantage for practical fast channel connection.

B. Comparison of Different Precoding Schemes

To first evaluate the efficiency of the proposed precoding schemes, we simulate the sum spectral efficiency with $K = 2$. As shown in Fig. 3 (b), for the FC transmitter, in the range of $\mathsf{SNR}_{\mathrm{BBF}} \leq 10\,\mathrm{dB}$ the simple BST scheme achieves the highest sum spectral efficiency, whereas when $\mathsf{SNR}_{\mathrm{BBF}} \gg 0\,\mathrm{dB}$ the BZF precoder performs better. Also, the curves of BZF precoders with different p values coincide with each other. Namely, the choice of p plays a trade-off between the beamwidth (i.e., the robustness to mobility) and the overhead for additional channel estimation. However, without severe blockages as in the given scenario, the choices of p does not change the sum spectral efficiency R_{sum}. Further, the OSPS transmitter achieves similar performance. As a comparison, we also simulate a recent hybrid precoding scheme proposed in [5] which is completely based on downlink channel reconstruction. As we can see, the proposed precoders achieve much better performance than that in [5].

C. Fully-Connected or One-Stream-Per-Subarray?

Note that the performance of different architectures highly depends on the channel condition ($\mathsf{SNR}_{\mathrm{BBF}}$) and the underlying precoders. A doubtless fact is that, the hardware complexity of Fig. 1 (b) is much lower than Fig. 1(a). For the same channel condition, Fig. 1 (b) requires a slightly less initial training overhead. As for the data communication phase, given the parameters in this paper, we can see from Fig. 4 (a) that by using the BST precoder under weak channel conditions (i.e., $\mathsf{SNR}_{\mathrm{BBF}} < 0\,\mathrm{dB}$) and using the BZF precoder under strong channel conditions (i.e., $\mathsf{SNR}_{\mathrm{BBF}} \geq 0\,\mathrm{dB}$), the two architectures achieve a similar sum spectral efficiency.

To evaluate the architecture power efficiency, otherwise stated, we consider the BST precoder. We first assume a reference scenario as the baseline, i.e, the OSPS architecture using the BST precoder and a SC modulation, with PAs of $P_{\mathrm{max},0} = 6$ dBm, $\eta_{\mathrm{max},0} = 0.3$. The backoff factor with respect to different waveforms and transmitter architectures can be written as $\alpha_{\mathrm{off}} = 1/(P_{\mathrm{PAPR}})$, where P_{PAPR} represents the PAPR of the input signals at the PAs. The investigation for 3GPP LTE in [11] showed that with a probability of 0.9999, the PAPR of the LTE SC waveform is smaller than $\sim 7.5\,\mathrm{dB}$ and the PAPR of the LTE orthogonal frequency division multiplexing (OFDM) waveform (with 512 subcarriers employing QPSK) is smaller than $\sim 12\,\mathrm{dB}$. We set P_{PAPR} to these values for Fig. 1 (b). In Fig. 1 (a), however, the input signals of the PAs are the sum of the signals from different RF chains. For OFDM signaling each signal can be modeled as a Gaussian random process [11] and the signals from different RF chains are independent, hence, the PAPR of the sum is the same as of one

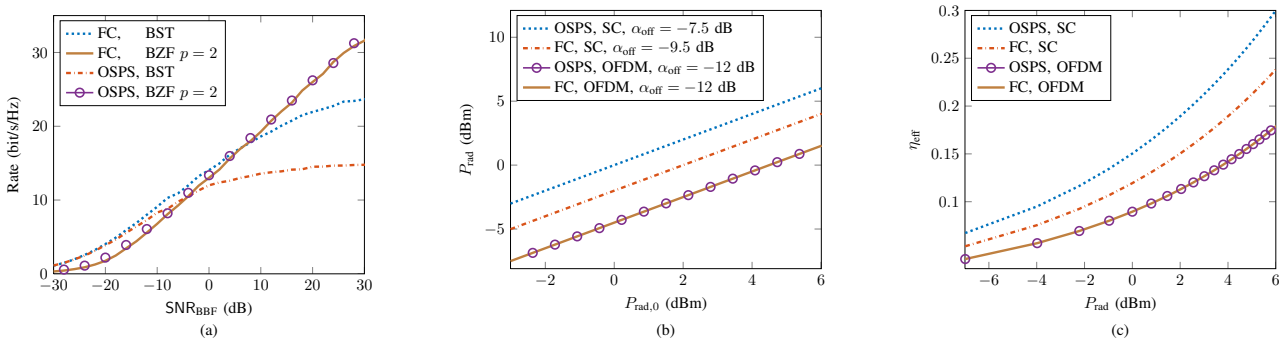

Fig. 4: The performance evaluation of different transmitter architectures in terms of (a) the sum spectral efficiency vs. increasing $\mathsf{SNR}_{\mathrm{BBF}}$, (b) the actual radiated power under *Option I* vs. the radiated power of the reference scenario, (c) the power efficiency under *Option II* vs. the actual radiated power.

RF chain. For the case of SC signaling there is no clear work in the literature that shows how the sum of SC signals behaves. We simulated the sum of $M_{\mathrm{RF}} = 2$ SC signals using the same parameters as in [11]. The result shows that with probability of 0.9999 the PAPR of the sum is smaller than $\sim 9.5\,\mathrm{dB}$. We apply these values and without loss of generality, we assume $\alpha_{\mathrm{off},0} = -7.5\,\mathrm{dB}$ for the reference scenario. As shown in (13), by deploying the same PAs (*Option I*), the two architectures achieve the same efficiency for a given P_{rad}. However as illustrated in Fig. 4 (b), given the same input signal (after the power compensation for the FC architecture) and precoding matrix as in the reference scenario, the OSPS architecture with SC signaling (OSPS, SC) achieves the highest P_{rad}, followed by (FC, SC), (OSPS, OFDM), and (FC, OFDM). In contrast, by deploying different PAs (*Option II*)[4], Fig. 4 (c) shows that (OSPS, SC) achieves the highest power efficiency, followed by (FC, SC), (OSPS, OFDM) and (FC, OFDM).

VI. Conclusion

In this paper, we proposed an analysis framework to evaluate the performance of typical hybrid transmitters at mmWave frequencies. In particular, we focused on the comparison of a fully-connected (FC) architecture and a one-stream-per-subarray (OSPS) architecture. We jointly evaluated the performance of the two architectures in terms of the initial beam alignment (BA), the data communication, and the transmitter power efficiency. We used our recently proposed BA scheme and a simple precoding scheme based on zero-forcing precoding of the effective channel after BA. Both schemes outperform the state-of-the-art counterparts in the literature and can be considered as the de-facto new state of the art. Given the parameters in this paper, our simulation results show that the two architectures achieve a similar sum spectral efficiency, but the OSPS architecture outperforms the FC case in terms of hardware complexity and power efficiency, only at the cost of a slightly longer time for the initial BA. We hope that the proposed work provides a good analysis framework for future mmWave MU-MIMO system design.

[4]Since $\eta_{\max}$ of different PAs highly depends on the technology, for simplicity, we assume that different PAs working in their linear range have roughly the same maximum efficiency $\eta_{\max,0}$.

References

[1] A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, S. L. H. Nguyen, L. Li, and K. Haneda, "Hybrid beamforming for massive MIMO: A survey," *IEEE Communications Magazine*, vol. 55, no. 9, pp. 134–141, 2017.

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

[3] J. Du, W. Xu, H. Shen, X. Dong, and C. Zhao, "Hybrid precoding architecture for massive multiuser MIMO with dissipation: Sub-connected or fully-connected structures?" *arXiv preprint arXiv:1806.02857*, 2018.

[4] P. L. Cao, T. J. Oechtering, and M. Skoglund, "Precoding design for massive MIMO systems with sub-connected architecture and per-antenna power constraints," in *WSA 2018; 22nd International ITG Workshop on Smart Antennas*, March 2018, pp. 1–6.

[5] M. R. Castellanos, V. Raghavan, J. H. Ryu, O. H. Koymen, J. Li, D. J. Love, and B. Peleato, "Channel-reconstruction-based hybrid precoding for millimeter-wave multi-user MIMO systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 2, pp. 383–398, 2018.

[6] X. Song, S. Haghighatshoar, and G. Caire, "Efficient beam alignment for mmWave single-carrier systems with hybrid MIMO transceivers," *arXiv preprint arXiv:1806.06425*, 2018.

[7] ——, "A scalable and statistically robust beam alignment technique for mm-Wave systems," *IEEE Trans. on Wireless Comm.*, vol. PP, pp. 1–1, 2018.

[8] N. N. Moghadam, G. Fodor, M. Bengtsson, and D. J. Love, "On the energy efficiency of MIMO hybrid beamforming for millimeter wave systems with nonlinear power amplifiers," *arXiv preprint arXiv:1806.01602*, 2018.

[9] V. Raghavan, S. Subramanian, J. Cezanne, A. Sampath, O. H. Koymen, and J. Li, "Single-user versus multi-User precoding for millimeter wave MIMO systems," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, pp. 1387–1401, June 2017.

[10] K. Venugopal, A. Alkhateeb, R. W. Heath, and N. G. Prelcic, "Time-domain channel estimation for wideband millimeter wave systems with hybrid architecture," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, Conference Proceedings, pp. 6493–6497.

[11] H. G. Myung, J. Lim, and D. J. Goodman, "Peak-to-average power ratio of single carrier FDMA signals with pulse shaping," in *Personal, Indoor and Mobile Radio Communications, 2006 IEEE 17th International Symposium on*. IEEE, Conference Proceedings, pp. 1–5.