Energy-Efficient Hybrid Precoding for mmWave Multi-user Systems

Vu Nguyen Ha[†], Duy H. N. Nguyen[‡], and Jean-Francois Frigon[†],
[†]École Polytechnique de Montréal, Poly-Grames Research Center, Montreal, Quebec, Canada, H3T 1J4
[‡]Department of Electrical and Computer Engineering, San Diego State University, San Diego, CA, USA 92182

***Abstract*—This paper aims to study an energy-efficiency (EE) maximization hybrid precoding (HP) design for mmWave multi-user (MU) systems where the analog precoding (AP) matrix is realized by a number of switches and phase shifters so that a connection between an RF chain and a transmit antenna can be switched off for energy saving. By explicitly considering the effect of each connection on the required power of digital precoding (DP) and AP design process, we describe the total power consumption as a sparsity form of the AP matrix. Together with the novel sparsity-modulus constraints of AP matrix, these sparsity terms make our system EE maximization (SEEM) problem be non-convex and challenging to solve. To tackle the SEEM problem, we first transform it into a subtractive-form weighted sum rate and power (WSRP) problem. We then exploit an alternating minimization of the mean-squared error algorithm to solve the WSRP problem where the DP vectors and AP matrix are updated alternatively, and a compressed sensing-based re-weighted quadratic-form relaxation method is employed to deal with the sparsity parts and the sparsity-modulus constraints.**

I. Introduction

Enabling spatial multiplexing with large antenna arrays can improve the spectral efficiency of mmWave systems [1], [2]. However, baseband precoding implementation for multi-stream transmissions can be highly complex with large antenna arrays. Thus, hybrid precoding (HP) has been considered as a promising alternative for mmWave MIMO systems [2]. This proposed transceiver architecture does not only reduce the number of RF chains by accomplishing a digital precoder and an analog precoder [2]–[4], [8]–[10] but also achieve the near-optimal performance thanks to the low-rank characteristics of mmWave channels [4]. However, the large size of the mmWave massive MIMO still raises a critical concern about the system energy efficiency (SEE), which is an important aspect of mmWave HP systems [5].

The study on SEE of mmWave HP MIMO systems have been presented in several recent papers [6], [7], [11], [12]. In particular, [6] studies the energy-efficient hybrid precoding (EE-HP) design for mmWave MIMO system with sub-arrays structure where the SEE is defined as ratio of sum-rate to the transmission power plus constant which is estimated based on the numbers of RF chains and antennas. Another EE-HP structure can be found in [7], where the number of RF chains can be optimized to economize the total power consumption. In [11], the authors propose a hybrid analog-digital architecture for HP in mmWave MIMO systems, where multiple sub-arrays are employed at the transmit and receive antennas. The EE-HP design problem for mmWave massive MIMO system is further studied in [12] where the numbers of RF chains and antennas are also optimized based on the statistical analysis values of EE for very large array antenna systems. To the best of our knowledge, the EE-HP design design for mmWave multiuser (MU) systems which optimizes the utilization of available RF chains, transmission antennas, and the connections among the RF chains and the antennas has been not studied in the literature.

Filling this gap, this paper studies a novel HP structure for mmWave MIMO system where each of the connections between the RF chains and antennas can be activated or inactivated (turned on or turned off) optimally to maximize the SEE. This new structure is then transferred into sparsity-modulus constraints of analog precoding (AP) matrix design. It also enables us to exploit the total power consumption as a sparsity function of AP matrix. We then formulate the SEE maximization (SEEM) problem as a new optimization problem, which is non-convex and challenging to solve. To tackle the SEEM problem, we first transform it into a subtractive-form weighted sum rate and power (WSRP) problem based on *"Dinkelbach method"* [13]. Subsequently, we exploit an alternating minimization of the mean-squared error (MMSE) algorithm to solve the WSRP problem where the DP vectors and AP matrix are updated alternatively. In each iteration, we employ re-weighted quadratic-form relaxation method to deal with the sparsity-modulus constraints and sparse formulation of total power consumption. In particular, this method helps us transform MMSE problem into a convex problem, which can be solved optimally. Extensive numerical studies are conducted where we examine the convergence and efficiency of the proposed algorithms as well as the impacts of different system parameters on the SEE.

II. System Model

A. Multi-user Hybrid Precoding System Model

Consider a downlink mmWave MU-HP scenario where a base station (BS) equipped with N_{T} antennas and N_{RF} RF chains serves K remote single-antenna users. Utilizing the HP technique, the BS first applies the DP vectors to the corresponding symbol sequences for the users. Specifically, a DP vector $\mathbf{w}_k \in \mathbb{C}^{N_{\mathrm{RF}}}$ is applied to the data symbol $s_k \in \mathbb{C}$, intended for user k. Without loss of generality, we assume $\mathbb{E}[|s_k|] = 1$. Following the digitally precoded sequences, the BS then employs an AP matrix, $\mathbf{A} \in \mathbb{C}^{N_{\mathrm{T}} \times N_{\mathrm{RF}}}$, to map the RF signals from N_{RF} RF chains to N_{T} antennas. In this work,

978-1-5386-3180-5/18/$31.00 ©2018 IEEE

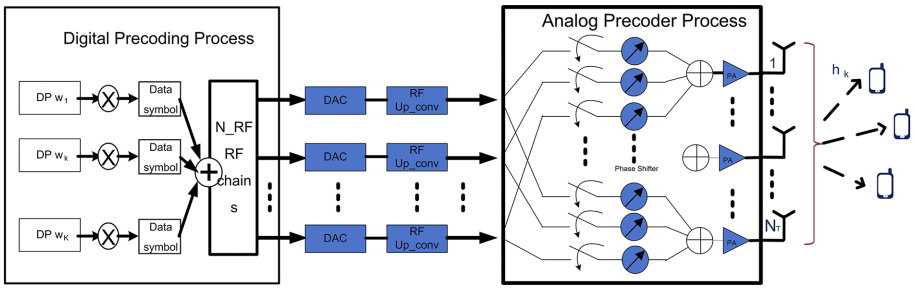

Fig. 1. Diagram of a mmWave MU system with hybrid analog/digital precoding design.

we consider a dynamic fully-connected RF chains to antennas structure in which $\mathbf{A}$ is implemented using sparsity-modulus analog phase shifter, i.e., modulus of each entry of $\mathbf{A}$ can be 1 or 0. Specifically, each connection between an RF chain and a specific antenna can be activated (turned on) or inactivated (turned off) flexibly, i.e., $\left|a_t^n\right|^2 = 1$ or 0 $\forall(t,n)$ where a_t^n represents the element on the row t and the column n of $\mathbf{A}$.

By taking into account of the HP design for MU system in [3], the estimated signal of user k can be given as

$$y_k = \mathbf{h}_k^H \sum_{\forall j} \mathbf{A}\mathbf{w}_j s_j + n_k, \tag{1}$$

where n_k is the additive Gaussian noise at user k and $\mathbf{h}_k \in \mathbb{C}^{N_\mathrm{T}}$ is the MISO channel from the BS to user k. Assuming coherent detection at the users, the signal-to-interference-plus-noise ratio (SINR) at user k can be given as

$$\mathrm{SINR}_k = \frac{\left|\mathbf{h}_k^H \mathbf{A}\mathbf{w}_k\right|^2}{\sum_{j\neq k}\left|\mathbf{h}_k^H \mathbf{A}\mathbf{w}_j\right|^2 + \sigma^2}, \tag{2}$$

where σ^2 is the power of additive Gaussian noise. Assuming Gaussian signalling between the BS and the users, the total achievable data-rate of the system is given by

$$R(\mathbf{W},\mathbf{A}) = \sum_{\forall k} \log(1+\mathrm{SINR}_k), \tag{3}$$

where $\mathbf{W}$ is denoted as the matrix generated by all DP vectors.

B. Energy Consumption Model

The total energy consumption in the system comprises of the energy consumed by all DP process, RF process, and transmission. For the DP process, the static power consumption at each user's signal is due to parts of baseband signal process. The energy consumption for this process can be considered unchanged due to our design which can be given as

$$P_\mathrm{DP} = K P_\mathrm{BB}, \tag{4}$$

where P_BB represents the power consumption of the baseband signal process. For the RF process, the baseband signal of each RF chain is first converted to the analog signal due to the real-time D/A processing and up-converted to the RF band due to the up-converter processing. Then, each RF signal is phase-shifted accordingly to the connection between that RF chain and a specific antenna. Afterwards, all signal heading to an antenna are mixed before going through the power amplifier process and being propagated by that antenna. Hence, an activated connection between an RF chain and an antenna can consume countable energy as illustrated in Fig. 1. Furthermore, one RF chain can be turned off for energy saving if there is no connection from that RF chain to all antennas. Likewise, an antenna may be inactive when all connections from RF chains to that antenna are turned off.

Denote P_DAC, P_RFC, P_PS, and P_PA as the power consumptions of D/A, RF converter, phase shifting, and power amplifier processes, respectively. It is worth to note that when the connection between RF chain n $(1 \le n \le N_\mathrm{RF})$ and antenna t $(1 \le t \le N_\mathrm{T})$ is activated, we have $\left|a_t^n\right|^2 = 1$, and $\left|a_t^n\right|^2 = 0$ implies that the corresponding connection is inactivated. Hence, we can express the total RF power consumption of RF process in a sparse form of $\mathbf{A}$ as follows:

$$\begin{aligned} P_\mathrm{RF}(\mathbf{A}) = {} & (P_\mathrm{DAC} + P_\mathrm{RFC})\|\mathbf{c}(\mathbf{A})\|_0 \\ & + P_\mathrm{PS}\|\mathbf{A}\|_0 + P_\mathrm{PA}\|\mathbf{r}(\mathbf{A})\|_0, \end{aligned} \tag{5}$$

where $\mathbf{c}(\mathbf{A}) = [c_1, ..., c_{N_\mathrm{RF}}]^T \in \mathbb{C}^{N_\mathrm{RF}}$ and $c_n = \sum_{\forall t}\left|a_t^n\right|^2$, $\mathbf{r}(\mathbf{A}) = [c_1, ..., c_{N_\mathrm{T}}]^T \in \mathbb{C}^{N_\mathrm{T}}$ and $r_t = \sum_{\forall n}\left|a_t^n\right|^2$. Next, the transmission power can be expressed as

$$P_\mathrm{T}(\mathbf{W},\mathbf{A}) = \sum_{\forall k} \mathbf{w}_k^H \mathbf{A}^H \mathbf{A}\mathbf{w}_k. \tag{6}$$

Therefore, the total energy consumption can be calculated as

$$P_\mathrm{tot}(\mathbf{W},\mathbf{A}) = P_\mathrm{DP} + P_\mathrm{RF}(\mathbf{A}) + P_\mathrm{T}(\mathbf{W},\mathbf{A}). \tag{7}$$

Then, the SEE (bits/Hz/W) is defined as the ratio of the achievable sum-rate and the total power consumption as follows:

$$\eta(\mathbf{W},\mathbf{A}) = \frac{R(\mathbf{W},\mathbf{A})}{P_\mathrm{tot}(\mathbf{W},\mathbf{A})}. \tag{8}$$

In this paper, we are interested in jointly optimizing the DP vectors and the AP matrix to maximize the SEE under the constraint on the transmission power budget of each antenna. In particular, this SEEM problem can be stated as

$$\max_{\mathbf{W},\mathbf{A}} \ \eta(\mathbf{W},\mathbf{A}) = \frac{\sum_{\forall k}\log(1+\mathrm{SINR}_k)}{P_\mathrm{DP} + P_\mathrm{RF}(\mathbf{A}) + P_\mathrm{T}(\mathbf{W},\mathbf{A})} \tag{9a}$$

$$\text{s. t.} \quad \left|a_t^n\right| = 1 \text{ or } 0, \forall(t,n), \tag{9b}$$

$$\sum_{\forall k} \mathbf{w}_k^H \mathbf{a}_t \mathbf{a}_t^H \mathbf{w}_k \le P_t^{\max}, 1 \le t \le N_\mathrm{T}, \tag{9c}$$

where $\mathbf{a}_t \in \mathbb{C}^{N_{\mathrm{Rf}}}$ is denoted as the vector corresponding to the t^{th} column of $\mathbf{A}^H$, and $P_t^{\max}$ is transmission power budget of antenna t at the BS.

III. Transformation of SEE Maximization Problem and General Design Algorithm

A. Transformation of SEE Maximization Problem

The objective of the SEEM problem (9) is in a fractional form, which is difficult to address. Therefore, we transform it into a subtractive form to aid the algorithm development. Toward this end, let us denote Θ as the set of feasible solutions $(\mathbf{W}, \mathbf{A})$ of SEEM problem, we can express the maximum SEE $\eta^\star$ as

$$\eta^\star = \eta(\mathbf{W}^\star, \mathbf{A}^\star) = \frac{R(\mathbf{W}^\star, \mathbf{A}^\star)}{P_{\mathrm{tot}}(\mathbf{W}^\star, \mathbf{A}^\star)} = \max_{(\mathbf{W},\mathbf{A})\in\Theta} \frac{R(\mathbf{W}, \mathbf{A})}{P_{\mathrm{tot}}(\mathbf{W}, \mathbf{A})}, \quad (10)$$

where $(\mathbf{W}^\star, \mathbf{A}^\star)$ represents the optimal solution. The following theorem provides the foundation based on which the transformation of problem (9) can be performed.

Theorem 1. *Let consider the following weighted sum-rate power (WSRP) maximization problem for given value of η as follows.*

$$\chi(\eta) = \max_{(\mathbf{W},\mathbf{A})\in\Theta} R(\mathbf{W}, \mathbf{A}) - \eta P_{\mathrm{tot}}(\mathbf{W}, \mathbf{A}). \quad (11)$$

Then, the maximum SEE $\eta^\star$ defined in (10) *is achieved if and only if*

$$\chi(\eta^\star) = \max_{(\mathbf{W},\mathbf{A})\in\Theta} R(\mathbf{W}, \mathbf{A}) - \eta^\star P_{\mathrm{tot}}(\mathbf{W}, \mathbf{A}) = 0, \quad (12)$$

for $R(\mathbf{W}, \mathbf{A}) \geq 0$ and $P_{\mathrm{tot}}(\mathbf{W}, \mathbf{A}) > 0$.

Proof: Theorem 1 can be proved by following the similar approach as in [13]. ∎

Since $R(\mathbf{W}^\star, \mathbf{A}^\star) - \eta^\star P_{\mathrm{tot}}(\mathbf{W}^\star, \mathbf{A}^\star) = 0$, $(\mathbf{W}^\star, \mathbf{A}^\star)$ is also an optimal solution of WSRP problem when $\eta = \eta^\star$; additionally, the optimal value of $\eta^\star \geq 0$ defined in (10) must satisfy $\chi(\eta^\star) = 0$. Therefore, to solve the complex SEEM problem, we can deal with the WSRP problem and adjusting the value of η. Specifically, we iteratively solve WSRP problem for a certain value of η and updating η until we reach the optimal $\eta^\star \geq 0$ satisfying $\chi(\eta^\star) = 0$.

B. General Design Algorithm

The solution of the SEEM problem can be addressed by updating η and solving the corresponding WSRP problem iteratively as given in Algorithm 1. In particular, we start with setting $\eta^{(0)} = 0$. In iteration n, after solving the WSRP problem corresponding to the updated value of $\eta^{(n)}$ to achieve the optimal value $(\mathbf{W}^\dagger, \mathbf{A}^\dagger)$, we check if $\chi(\eta^{(n)})$ is small enough or not with a tolerance ϵ. The process stops if $\chi(\eta^{(n)}) \leq \epsilon$. Otherwise, the process continues by updating $\eta^{(n+1)} = \frac{R(\mathbf{W}^\dagger, \mathbf{A}^\dagger)}{P_{\mathrm{tot}}(\mathbf{W}^\dagger, \mathbf{A}^\dagger)}$. Here, we can see that the value of η in the next iteration increases if $\chi(\eta^{(n)}) > 0$ and vice versa.

Algorithm 1 General Design Algorithm

1: Initialize arbitrary Maximum iteration number N, tolerance ϵ, $\eta^{(0)} = 0$, FLAG = **false**, and $n = 1$.
2: **repeat**
3: Solve the problem (11) with the recent value of $\eta^{(n)}$ to achieve $(\mathbf{W}^{(n)}, \mathbf{A}^{(n)})$.
4: **if** $|R(\mathbf{W}^{(n)}, \mathbf{A}^{(n)}) - \eta^{(n)} P_{\mathrm{tot}}(\mathbf{W}^{(n)}, \mathbf{A}^{(n)})| \leq \epsilon$ **then**
5: Set FLAG = **true**.
6: **else**
7: Set $\eta^{(n+1)} = \frac{R(\mathbf{W}^{(n)}, \mathbf{A}^{(n)})}{P_{\mathrm{tot}}(\mathbf{W}^{(n)}, \mathbf{A}^{(n)})}$.
8: Update $n := n + 1$.
9: **end if**
10: **until** FLAG = **true** or $n > N$.
11: Return $(\mathbf{W}^{(n)}, \mathbf{A}^{(n)})$ if $n \leq N$, or $(\mathbf{W}^{(n-1)}, \mathbf{A}^{(n-1)})$ if $n > N$.

IV. Hybrid Precoding for Solving WSRP Problem

In this section, we consider the WSRP problem, which is NP-hard as a result of the non-convex sum rate, the ℓ_0-norm of matrix $\mathbf{A}$ in the objective function, and the sparsity-modulus constraint (9b). Hence, finding its globally optimal solution is prohibitively complex. In this case, an efficient (probably suboptimal) solution is more sought instead. In what follows, we will provide such an efficient solution which employs the compressed sensing method.

A. Sparsity Relaxation

It is worth noting that the ℓ_0-norms of matrix $\mathbf{A}$, $\mathbf{r}(\mathbf{A})$, and $\mathbf{c}(\mathbf{A})$, can be defined directly from $\|a_t^n\|_0$. Interestingly, the sparsity solutions can be obtained by employing the re-weighted ℓ_1-norm minimization methods, originally proposed to enhance the data acquisition in compressed sensing. In particular, the method applies a weighting factor on each element of the matrix or vector corresponding to the sparsity part of the optimization problem. Additionally, these weighting factors should be chosen and updated in each iteration such that the solutions corresponding to elements with large values have to be penalized.

Due to the fact that $|a_t^n| = 1$ or 0, a good choice of the weighting factors corresponding to a_t^n, c_n, and r_t in iteration $(n + 1)$ can be chosen as [15]

$$\psi_t^n = \frac{1}{|a_t^n|^2 + \varepsilon}, \phi_n = \frac{1}{\sum_{\forall t} |a_t^n|^2 + \varepsilon}, \varphi_t = \frac{1}{\sum_{\forall n} |a_t^n|^2 + \varepsilon}, \quad (13)$$

where $\varepsilon \ll 1$. Based on these weighting factors, we can approximate the ℓ_0-norm terms in the objective function of WSRP problem to indicator functions in the quadratic form as follows:

$$\|\mathbf{A}\|_0 = \sum_{t=1}^{N_{\mathrm{T}}} \|\mathbf{a}_t\|_0 = \sum_{t=1}^{N_{\mathrm{T}}} \mathbf{a}_t^H \Psi_t \mathbf{a}_t, \quad (14)$$

$$\|\mathbf{c}(\mathbf{A})\|_0 = \sum_{t=1}^{N_{\mathrm{T}}} \sum_{n=1}^{N_{\mathrm{RF}}} \phi_n \mathbf{a}_t^H \mathbf{F}_t \mathbf{a}_t, \quad (15)$$

$$\|\mathbf{r}(\mathbf{A})\|_0 = \sum_{t=1}^{N_{\rm T}} \varphi_t \mathbf{a}_t^H \mathbf{I} \mathbf{a}_t, \tag{16}$$

where $\Psi_t = diag(\psi_t^1, ..., \psi_t^{N_{\rm RF}})$, $\mathbf{F}_t = diag(0_{1\times(t-1)}, 1, 0_{1\times(N_{\rm RF}-t)})$, and $\mathbf{I}$ is the identity matrix with the size of $N_{\rm RF} \times N_{\rm RF}$. Then, the total RF power consumption can be approximated to the convex quadratic-form as

$$P_{\rm RF} = \sum_{t=1}^{N_{\rm T}} \mathbf{a}_t^H \mathbf{D}_t \mathbf{a}_t, \tag{17}$$

where

$$\mathbf{D}_t = P_{\rm PS}\Psi_t + P_{\rm PA}\varphi_t \mathbf{I} + (P_{\rm DAC} + P_{\rm RFC})\mathbf{F}_t \sum_{\forall n} \phi_n \tag{18}$$

By properly choosing and updating Ψ_t's, ϕ_n's, and φ_t's iteratively, the non-convex dis-continuous ℓ_0-norms can be effectively approximate to the quadratic forms. Then, the WSRP problem can be approximated to the following problem.

$$\begin{aligned} \max_{\mathbf{W},\mathbf{A}} \; & l(\mathbf{A},\mathbf{W}) = \sum_{\forall k} \log(1+\mathrm{SINR}_k) \\ & -\eta\Big[P_{\rm DP} + \sum_{t=1}^{N_{\rm T}} \mathbf{a}_t^H \mathbf{D}_t \mathbf{a}_t + \sum_{t=1}^{N_{\rm T}} \mathbf{a}_t^H \big(\sum_{\forall k} \mathbf{w}_k \mathbf{w}_k^H\big)\mathbf{a}_t\Big] \\ \text{s. t.} \quad & \text{contraints (9b) and (9c).} \end{aligned} \tag{19}$$

B. MMSE-based Transformation

In this section, we address the non-convex problem (19) by relating it to a weighted sum-mean square error (MSE) minimization problem as mentioned in the following Theorem.

Theorem 2. *The problem (19) is equivalent to the following weighted sum-MSE minimization problem*

$$\begin{aligned} \min_{\mathbf{A},\mathbf{W},\{\delta_k,\omega_k\}} \; & g(\mathbf{W},\mathbf{A},\delta_k,\omega_k) = \sum_{\forall k}\Big(\omega_k \mathbb{E}\Big[\big|s_k - \delta_k y_k\big|^2\Big] - \log\omega_k - 1\Big) \\ & +\eta\Big[P_{\rm DP} + \sum_{t=1}^{N_{\rm T}} \mathbf{a}_t^H \big(\mathbf{D}_t + \sum_{\forall k} \mathbf{w}_k \mathbf{w}_k^H\big)\mathbf{a}_t\Big] \\ \text{s.t.} \quad & \textit{contraints} \text{ (9b) } \textit{and} \text{ (9c).} \end{aligned} \tag{20}$$

where ω_k and δ_k represent the MSE weight and the receive coefficient for user k, respectively.

Proof: The proof is given in Appendix A. ∎

It is noted that the objective function in problem (20) is not *jointly* convex, but it is convex over each set of variables $\mathbf{w}_k$'s, $\mathbf{a}_t$'s, δ_k's, and ω_k's. Hence, an efficient algorithm for solving this problem can be developed based on the basic idea of alternating between WMMSE optimization of $\mathbf{w}_k$'s and $\mathbf{a}_t$'s, and the MSE weight update for δ_k's, and ω_k's.

C. Update MSE weights and receive coefficients

For given $(\mathbf{W}, \mathbf{A})$, δ_k's, and ω_k's can be determined according to the results in Appendix A. In particular, the MMSE receive filter at user k is given as

$$\delta_k^\star = \delta_k^{\rm MMSE} = \Big(\sum_{\forall j} |\mathbf{h}_k^H \mathbf{A}\mathbf{w}_j|^2 + \sigma_k^2\Big)^{-1} \mathbf{w}_k^H \mathbf{A}^H \mathbf{h}_k. \tag{21}$$

And, the optimum value of ω_k can be expressed as

$$\omega_k^\star = e_k^{-1} = 1 + \Big(\sum_{\forall j} |\mathbf{h}_k^H \mathbf{A}\mathbf{w}_j|^2 + \sigma_k^2\Big)^{-1} |\mathbf{w}_k^H \mathbf{A}^H \mathbf{h}_k|^2. \tag{22}$$

D. Digital Precoding Design

For given AP matrix $\mathbf{A}$, the optimal $\mathbf{w}_k$'s can be obtained by solving the following Quadratically Constrained Quadratic Program (QCQP).

$$\begin{aligned} \min_{\{\mathbf{w}_k\}} \; & \sum_{\forall(k)} \Big[\mathbf{w}_k^H \Big(\sum_{\forall j} \omega_j |\delta_j|^2 \mathbf{A}^H \mathbf{h}_j \mathbf{h}_j^H \mathbf{A} + \eta \mathbf{A}^H \mathbf{A}\Big)\mathbf{w}_k \\ & \qquad -\omega_j \delta_k' \mathbf{w}_k^H \mathbf{A}^H \mathbf{h}_k - \omega_j \delta_k \mathbf{h}_k^H \mathbf{A} \mathbf{w}_k\Big] \\ \text{s.t.} \quad & \text{constraint (9c).} \end{aligned} \tag{23}$$

This such QCQP problem can be solved by any standard convex optimization solvers or the Lagrangian duality method.

E. Sparsity-modulus Analog Phase-Shifting Matrix Design

For given DP vectors $\mathbf{w}_k$'s, the AP matrix $\mathbf{A}$ can be achieved by solving the following problem.

$$\begin{aligned} \min_{\{\mathbf{a}_t\}} \; & \sum_{t=1}^{N_{\rm T}} \Big(\mathbf{a}_t^H \Lambda_t \mathbf{a}_t + \eta \mathbf{a}_t^H \Omega_{\mathbf{w}} \mathbf{a}_t + \eta \mathbf{a}_t^H \mathbf{D}_t \mathbf{a}_t \\ & \qquad - \sum_{\forall k} \omega_k \delta_k' h_{k,t} \mathbf{w}_k^H \mathbf{a}_t - \mathbf{a}_t^H \sum_{\forall k} \omega_k \delta_k h_{k,t}' \mathbf{w}_k\Big) \\ \text{s.t.} \quad & \text{constraints (9b) and (9c).} \end{aligned} \tag{24}$$

where $\Lambda_t = \sum_{\forall k} \omega_k |\delta_k|^2 |h_{k,t}|^2 \Omega_{\mathbf{w}}$ and $\Omega_{\mathbf{w}} = \sum_{\forall k} \mathbf{w}_k \mathbf{w}_k^H$. The difficulty of solving the problem (24) mainly comes from the sparsity-modulus constraint (9b). It has been shown in [15] that applying the re-weight method with updating the weight matrix $\mathbf{D}_t$ as in (13) will let $\mathbf{a}_t^H \mathbf{D}_t \mathbf{a}_t$ approximate to the ℓ_0-norm term $P_{\rm RF}$ after a certain number of iterations. However, this method cannot ensure the conformity of the sparsity-modulus constraint (9b). To overcome this challenge, we employ the compressed sensing method [15] once again through the following result.

Lemma 1. *If $|a_t^n| = 0$ or 1, then we can approximate the vector $\mathbf{a}_t$ as follows*

$$\mathbf{a}_t \approx \Psi_t^{1/2} \mathbf{a}_t \approx \Psi_t^{-1/2} \mathbf{a}_t. \tag{25}$$

Proof: Let $\mathbf{a}_t' = \Psi_t^{1/2}\mathbf{a}_t$ and $\mathbf{a}_t'' = \Psi_t^{1/2}\mathbf{a}_t$. Then, we have $a_t^{n\prime} = a_t^n/\sqrt{\psi_t^n}$ and $a_t^{n\prime\prime} = a_t^n\sqrt{\psi_t^n}$. If $|a_t^n| = 0$, we have $a_t^{n\prime} = a_t^{n\prime\prime} = a_t^n = 0$. Otherwise, when $|a_t^n| = 1$, we can conclude that $\sqrt{\psi_t^n} \approx 1$; hence, we also have $a_t^{n\prime} \approx a_t^n/\sqrt{\psi_t^n}$ and $a_t^{n\prime\prime} \approx \sqrt{\psi_t^n} a_t^n$. Finally, we have $\mathbf{a}_t \approx \mathbf{a}_t' = \Psi_t^{1/2}\mathbf{a}_t$ and $\mathbf{a}_t \approx \mathbf{a}_t'' = \Psi_t^{-1/2}\mathbf{a}_t$. ∎

Lemma 1 encourages us to replace $\mathbf{a}_t$ in problem (24) by the term $\Psi_t^{1/2}\mathbf{a}_t$ or $\Psi_t^{-1/2}\mathbf{a}_t$ and also relax the the sparsity-modulus constraints (9b) to the convex constraints, i.e., $|a_t^n| \le 1$ for all (t, n). Then, the sparsity-solution can be achieved by applying the transformation in (25) one more time. In addition, it is worth to note that the problem (24) can be decomposed into $N_{\rm T}$ sub-problems corresponding to $N_{\rm T}$ vectors $\mathbf{a}_t$'s to

Algorithm 2 ITERATIVE WEIGHTED RATE AND POWER MAXIMIZATION HYBRID BEAMFORMING

1: Initialize by setting $\mathbf{w}_k^{(0)} = \theta \mathbf{1}_{N_{\mathrm{RF}} \times 1}$ and $\mathbf{a}_t = \mathbf{1}_{N_{\mathrm{RF}} \times 1}$ for all (k,t), where θ (> 0) is small enough to satisfy the power constraint, and $l = 0$.
2: **repeat**
3: Update $\Psi_t^{(l)}$'s, $\phi_n^{(l)}$'s, $\varphi_t^{(l)}$'s as in (13).
4: Calculate $\delta_k^{(l+1)}$'s as in (21).
5: Calculate and $\omega_k^{(l+1)}$'s as in and (22).
6: Determine $\mathbf{w}_k^{(l+1)}$'s by solving problem (23) corresponding to $\mathbf{A}^{(l)}$'s, $\delta_k^{(l)}$'s and $\omega_k^{(l)}$'s.
7: **for** $t = 1$ to N_{T} **do**
8: Solve problem (26) corresponding to $\mathbf{W}^{(l+1)}$'s, $\delta_k^{(l)}$'s and $\omega_k^{(l)}$'s, to determine $\mathbf{a}_t^{(l+1)}$.
9: **end for**
10: Return $\mathbf{A}^{(l+1)}$ from $\mathbf{a}_t^{(l+1)}$'s.
11: Update $l := l + 1$.
12: **until** Convergence or a stopping criterion trigger.
13: Return $(\mathbf{W}^{(l)}, \mathbf{A}^{(l)})$.

reduce the complexity of solving. Specifically, the sub-problem corresponding to $\mathbf{a}_t$ can be stated as follows:

$$(\mathcal{P}_t) \ \min_{\mathbf{a}_t} \ \mathbf{a}_t^H \Phi_t \mathbf{a}_t - \mathbf{f}_t^H \mathbf{a}_t - \mathbf{a}_t^H \mathbf{f}_t \tag{26a}$$

$$\text{s.t.} \ \ |a_t^n| \leq 1, \ \ \forall n, \tag{26b}$$

$$\mathbf{a}_t^H \Psi_t^{1/2} \Omega_{\mathbf{w}} \Psi_t^{1/2} \mathbf{a}_t \leq P_t^{\max}, \tag{26c}$$

where $\Phi_t = \Psi_t^{1/2}\big(\Lambda_t + \eta\Omega_{\mathbf{w}} + \eta \mathbf{D}_t\big)\Psi_t^{1/2}$ and $\mathbf{f}_t = \Psi_t^{-1/2} \sum_{\forall k} \omega_k \delta_k h'^{H}_{k,t} \mathbf{w}_k^H$. Again, the problem $(\mathcal{P}_t)$ is a QCQP-form problem which can be solved by employing any standard optimization solver.

By iteratively updating $\{\mathbf{w}_k, \mathbf{a}_t, \delta_k, \omega_k\}$, we can obtain the MMSE hybrid precoding. Combined with the compressed sensing-based method with updating the weight factors Ψ_t's, φ_t's, and ϕ_n's, the DP vectors and sparsity-modulus AP matrix design for WSRP maximization is summarized in Algorithm 2. Giving the convergence proof of this algorithm is challenging due to the lack of space. However, this algorithm relies on the iterative weighted updates in (13) and (25). Hence, it can be considered as a special case of Majorization-Minimization algorithms [15] whose convergence proof is given.

V. SIMULATION RESULTS

We consider a MISO system where a 6×6 uniform polar array (UPA - $M = 36$) with antenna spacing equal to a half-wavelength is adopted at the BS. The channel to each user contains four clusters of 10 paths, i.e., $C = 4$, $L = 10$. All the channel path gains $\alpha_{c,\ell}$'s are assumed to be i.i.d. Gaussian random variables with variance σ_α^2. The azimuth angles are assumed to be uniformly distributed in $[0; 2\pi]$, and the AoA/AoD elevation angles are uniformly distributed in $[-\frac{\pi}{2}; \frac{\pi}{2}]$. The noise variance σ^2 is set at 10^{-13}. In this simulation, we employ the path loss ABG model for macro-cellular scenario with $f = 60\ GHz$ in Table I of [16], where

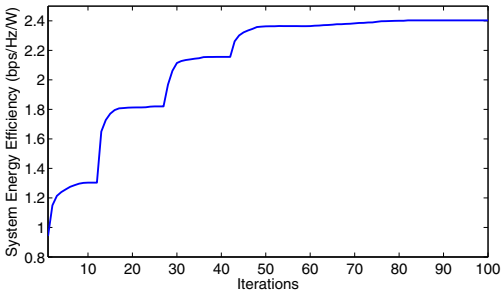

Fig. 2. Convergence of proposed algorithms.

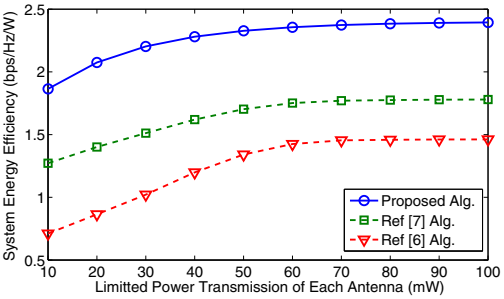

Fig. 3. System energy efficiency *versus* $P_t^{\max}$ where $K = N_{\mathrm{RF}} = 8$.

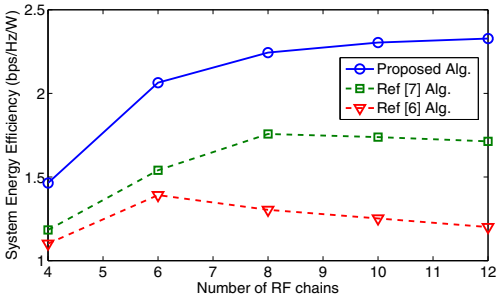

Fig. 4. System energy efficiency *versus* N_{RF} where $K = 4$.

distance range is set equal to 100 m for all users. In addition, we set $P_t^{\max} = 50\ mW$ (except Fig. 3), $P_{\mathrm{BB}} = 300\ mW$, $P_{\mathrm{DAC}} = 200\ mW$, $P_{\mathrm{RFC}} = 43\ mW$, $P_{\mathrm{PS}} = 30\ mW$, and $P_{\mathrm{PA}} = 20\ mW$ [17].

We illustrate the convergence of our proposed algorithm (Algorithm 1 integrated with Algorithm 2) in Fig. 2 where the variations of SEE over the iterations are shown. To obtain results in this figure, both K and N_{RF} are set equal to 8. As can be seen, the SEE, which is calculated based on the outcome $(\mathbf{W}, \mathbf{A})$ of Algorithm 2 in each iteration, increases monotonically before reaching its the maximum value which also is the outcome of Algorithm 1.

In Fig.s 3, 4, and 5 present the SEE achieved by our proposed algorithm versus the transmission-power budget of each antenna ($P_t^{\max}$), the number of RF chains (N_{RF}), and the number of users (K), respectively. In addition, the results achieved by two EE-HP algorithms in [6] and [7] are also

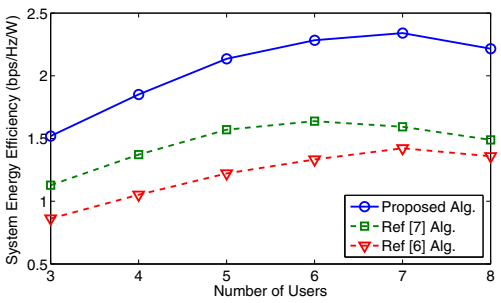

Fig. 5. System energy efficiency *versus* the number of users where $N_{\mathrm{RF}} = 8$.

illustrated for the comparison purpose. As observed from the figures, the achievable SEEs achieved by our proposed algorithm are much higher than that achieved by the two methods given in [6] and [7] in all schemes. In Fig. 3, the SEEs achieved by all algorithms increase as $P_t^{\max}$ increases before saturating at the high regime of $P_t^{\max}$. In Fig. 4, our proposed algorithm can enhance more SEE when N_{RF} enlarges. This is because the larger value of N_{RF} results in the higher freedom for designing. However, the SEE due to the other two methods descend as N_{RF} becomes large. It again confirms the superior performance of our proposed design. Interestingly, Fig. 5 illustrates that the SEE achieved by all schemes increases and then decreases with the number of users. This can be explained as the power consumption of baseband signal processing is very high to support more users. Furthermore, the increase rate in the required power is much faster than the rate in the sum-rate improvements.

VI. Conclusion

This paper has considered the dynamic fully-connected AP structure for the HP mmWave MU systems where each connection between one RF chain and one antenna can be activated or inactivated brilliantly. This dynamic fully-connected AP structure has been represented as a novel sparsity-modulus constraint for the AP matrix design. In addition, the total power consumption of this system is formulated generally as the sparsity form of AP matrix. Then, a new compressed sensing-based EE-HP algorithm has been proposed to maximize the SEE. Numerical results have illustrated that our proposed algorithms outperform the reference ones in literature and confirmed the efficiency of our proposed algorithms.

Appendix A
Proof of Theorem 2

We notice that $R(\mathbf{W}, \mathbf{A})$ can be expressed as a function of the error covariance matrix after MMSE receive filtering [14]. The MMSE receive filter at user k is given as

$$\begin{aligned}\delta_k^{\mathrm{MMSE}} &= \arg\min_{\delta_k} \mathbb{E}\{\, |s_k - \delta_k y_k|^2 \,\} \\ &= \left(\sum_{\forall j} |\mathbf{h}_k^H \mathbf{A}\mathbf{w}_j|^2 + \sigma_k^2\right)^{-1} \mathbf{w}_k^H \mathbf{A}^H \mathbf{h}_k. \quad (27)\end{aligned}$$

The MSE-value for user k given that the MMSE-receive filter is applied can be written as $\varepsilon_k = \mathbb{E}\{\, |s_k - \delta_k^{\mathrm{MMSE}} y_k|^2 \,\} = \left[1 + \mathrm{SINR}_k\right]^{-1}$. Hence, we have

$$R_k(\mathbf{A}, \mathbf{W}) = \log(e_k^{-1}) = -\log(e_k) \quad (28)$$

In addition, it is easy to see that the optimum value of δ_k and ω_k can be expressed as

$$\delta_k^\star = \delta_k^{\mathrm{MMSE}} \text{ and } \omega_k^\star = e_k^{-1}. \quad (29)$$

At that point we have $g(\mathbf{A}, \mathbf{W}, \delta_k, \omega_k) = l(\mathbf{A}, \mathbf{W})$, then we have finished the proof.

Acknowledgement

This work was supported in part by the National Sciences and Engineering Research Council of Canada under Grant RGPIN-2016-06401 and in part by a startup fund from San Diego State University.

References

[1] J. Andrews,et. al., "What will 5G be?" *IEEE J. Select. Areas in Commun.,* vol. 32, no. 6, pp. 1065–1082, Jun. 2014.
[2] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.,* vol. 13, no. 3, pp. 1499–1513, Mar. 2014.
[3] A. Alkhateeb, G. Leus, and R. W. Heath, "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE Trans. Wireless Commun.,* vol. 14, no. 11, pp. 6481–6494, Nov. 2015.
[4] T. E. Bogale, L. B. Le, A. Haghighat and L. Vandendorpe, "On the number of RF chains and phase shifters, and scheduling design with hybrid analog-digital beamforming," *IEEE Trans. Wireless Commun.,* vol. 15, no. 5, pp. 3311–3326, May 2016.
[5] K. N. R. S. V. Prasad, E. Hossain, and V. K. Bhargava, "Energy efficiency in massive mimo-based 5g network: Opportunities and challenges," *IEEE Wireless Commun.,* vol. 24, no. 3, pp. 86–94, June 2017.
[6] X. Gao, et. al., "Energy-efficiency hybrid analog and digital precoding for mmwave mimo systems with large antenna arrays," *IEEE J. Sel. Areas Commun.,* vol. 34, no. 4, pp. 998–1009, Apr. 2016.
[7] S. He, J. Wang, Y. Huang, B. Ottersten, and W. Hong, "Codebook-based hybrid precoding for millimeter wave multiuser systems," *IEEE Trans. Signal Process.,* vol. 65, no. 20, pp. 5289–5304, Oct. 2017.
[8] D. H. N. Nguyen, L. B. Le, T. Le-Ngoc, R. W. Heath Jr., "Hybrid MMSE precoding and combining designs for mmWave Multiuser systems," *IEEE Access*, vol. 5, pp. 19167–19181, Sept. 2017.
[9] V. N. Ha, D. H. N. Nguyen, and J.-F. Frigon, "Joint subchannel allocation and hybrid precoding design for mmWave multiuser ofdma systems," *IEEE PIMRC 2017,* Oct. 2017.
[10] V. N. Ha, D. H. N. Nguyen, and J.-F. Frigon, "Subchannel allocation and hybrid precoding in mmWave ofdma systems," submitted to *IEEE Trans. Wireless Commun.,* 2017.
[11] D. Zhang, Y. Wang, X. Li, and W. Xiang, "Hybridly-connected structure for hybrid beamforming in mmwave mimo systems," *IEEE Trans. Commun.,* vol. PP, no. 99, 2017.
[12] R. Zi, X. Ge, J. Thompson, C.-X. Wang, H. Wang, and T. Han, "Energy efficiency optimization of 5g radio frequency chain systems," *IEEE J. Sel. Areas Commun.,* vol. 34, no. 4, pp. 758–771, Apr. 2016.
[13] W. Dinkelbach, "On nonlinear fractional programming," *Bulletin of the Australian Mathematical Society,* vol. 13, pp. 492–498, Mar. 1967.
[14] S. S. Christensen, et. al., "Weighted sum-rate maximization using weighted MMSE for MIMO-BC beamforming design," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 4792–4799, Dec. 2008.
[15] E. Candes, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Analysis Applications,* vol. 14, no. 5, pp. 877–905, Dec. 2008.
[16] S. Sun, et. al., "Propagation path loss models for 5G urban micro- and macro-cellular scenarios," in proc. *IEEE VTC2016-Spring,* May 2016.
[17] W. Li, et. al., "60-GHz 5-bit phase shifter with integrated VGA phase-error compensation," *IEEE Trans. Microw. Theory Techn.,* vol. 61, no. 3, pp. 1224–1235, Mar. 2013.