Sum Rate Analysis of Massive MIMO Downlink With Hybrid Beamforming

Dhanushka Kudathanthirige and Gayan Amarasuriya
Department of Electrical and Computer Engineering, Southern Illinois University, Carbondale, IL, USA 62901
Email: {dhanushka.kudathanthirige, gayan.baduge}@siu.edu

Abstract—The achievable sum rate of multi-cell multi-user massive multiple-input multiple-output (MIMO) downlink is investigated. Hybrid beamformers, which are constructed by using estimated/imperfect channel state information, are employed at the base-station. Thus, the massive MIMO base-station in each of the \( L \) cells is enabled with a typical power-intensive digital precoder and a quantized analog beamformer. This set-up can significantly reduce the number of radio frequency (RF) chains required for the digital precoder, and thereby, the power consumption in the digital signal processing and the complexity of circuitry at the massive MIMO base-station. The achievable asymptotic sum rate expressions are derived for linear digital precoders namely, (i) zero-forcing transmission and (ii) maximum ratio transmission cascaded with a phase-shifting analog beamformer. Thereby, the asymptotic sum rate degradation due to the hybrid beamforming is quantified and compared against the full-dimensional digital beamforming. This sum rate loss is a function of the number of phase quantization levels and cannot be canceled completely even in the asymptotic base-station antenna regime. Nevertheless, our analysis reveals that the detrimental effects of phase quantization and reduced number of RF chains of the hybrid beamforming can be mitigated in the limit of increasingly many base-station antennas when the receiver thermal noise power is negligibly smaller than the residual interference due to pilot contamination.

I. INTRODUCTION

Massive multiple-input multiple-output (MIMO) base-stations are capable of utilizing aggressive spatial multiplexing techniques and efficient power scaling laws to harness unprecedented spectral and energy efficiency gains by exploiting large antenna arrays [1]. Moreover, the excess degrees-of-freedom provided by these large antenna arrays can be used for wireless energy harvesting/transfer applications, provisioning physical-layer security, and enabling full-duplex transmission schemes. However, these performance gains are provided at the cost of full-dimensional digital signal processing (DSP) by means of spatial multiplexing, which utilizes a dedicated radio frequency (RF) chain for each antenna at the base-station. In this context, having dedicated RF chains arises several challenges in terms of increased cost, power consumption in DSP and complexity of circuitry at the base-station when the number of antennas grows without bound. Nevertheless, these challenges can be addressed by introducing a hybrid beamforming structure [2], which consists of a digital baseband precoder and an analog RF beamformer.

Some of the notable research contributions to the growth of massive MIMO systems can be outlined as follows: To begin with, reference [1] investigates non-cooperative cellular multi-user massive MIMO systems and concludes that the asymptotic performance is only limited by the pilot contamination, which arises from the residual interference due to reuse of non-orthogonal pilot sequences. Transmit power scaling laws for the user nodes are investigated in [3] for uplink transmissions of multi-user massive MIMO systems by considering the perfect and imperfect channel state information (CSI). Thereby, in [3], the asymptotic spectral and energy efficiency expressions are derived in closed-form, and rate-energy trade-off is quantified. In [4], the uplink channel estimation techniques based on the minimum mean square error (MMSE) criterion are studied with the effect of antenna correlation, and thereby, the asymptotic performance metrics of massive MIMO systems are derived by using the deterministic-equivalence approach. Furthermore, in [4], the number of antennas required to achieve a predefined percentage of the asymptotic performance limits is computed.

In hybrid beamforming architecture, a high-dimensional RF beamformer is cascaded with a low-dimensional baseband DSP unit through limited number of RF chains, which greatly reduces the complexity of the design [2], [5], [6]. Several important contributions to the development of hybrid beamforming techniques for massive MIMO can be summarized as follows: In [2], a low-dimensional zero-forcing (ZF) precoder, which is implemented by using the equivalent channel obtained via the product of RF beamformer and the true channel matrix, is investigated for the case of perfect CSI. Reference [5] investigates the achievable rate performance of analog-digital combining in multi-user massive MIMO with maximum ratio transmission (MRT) and ZF precoding in the presence of genie-aided perfect CSI. Furthermore, in [5], it is shown that the sum rate degradation due to hybrid beamforming can be compensated through increasing the number of antennas by 27\%. Reference [6] and [7] investigate the adaptive hybrid precoder designs. In [6], the RF and baseband precoders can adapt to mitigate the inter-cell and inter-pair interference, respectively. Moreover, [7] proposes an additional adaptive block, which determines the non-zero elements of RF precoder. The aforementioned references [2]-[7] consider hybrid precoding for multi-user single-cell massive MIMO systems with perfect CSI. Moreover, reference [8] investigates precoder designs to improve weighted sum rate and max-min fairness in cooperative multi-user massive MIMO systems. In [9], the impact of channel estimation on the achievable rates of massive MIMO system with codebook-based hybrid beamforming is studied.

In this paper, we investigate the asymptotic achievable sum rate of multi-cell multi-user massive MIMO downlink with cascaded digital precoding (MRT and ZF) and RF/analog beamforming designs based on estimated/imperfect CSI. To this end, each massive MIMO base-station in the \( L \) cell system...
is equipped with a low-dimensional baseband digital precoder and a high-dimensional analog RF beamformer, which has a constant amplitude and quantized phase. The elements of both precoder matrices are designed by utilizing the imperfect CSI estimated through the uplink pilots transmitted by the user nodes. In this context, the uplink channels are estimated at the base-station by using shared orthogonal pilot sequences transmitted by user nodes in $L$ cells, and consequently, the effects of inter-cell pilot contamination are taken into account for the performance analysis. Then, the asymptotic sum rate expressions are derived in closed-form for MRT and ZF precoding when the number of antennas at the base-station grows unbounded. Our analysis reveals that the hybrid beamforming is subjected to a trade-off between the achievable sum rate and the complexity of DSP. Nevertheless, when the residual interference due to pilot contamination is significantly higher than the noise power at the user nodes, the unfavorable effects of hybrid beamforming with imperfect CSI can be asymptotically mitigated in the limit of infinitely many base-station antennas.

**Notation:** $\mathbf{Z}^H$, $\text{Tr}(\mathbf{Z})$, $||\mathbf{Z}||$ and $[\mathbf{Z}]_{k,l}$ denote the Hermitian-transpose, trace, Euclidean norm and the $(k,l)$th element of the matrix $\mathbf{Z}$, respectively. The $\mathbb{E}[]$ and $\mathbb{V}ar[\cdot]$ are the expectation and variance operators, respectively. Furthermore, $\mathbf{Z} \sim R(\theta)$ denotes that $\mathbf{Z}$ is Rayleigh distributed with parameter $\theta$, and $\mathbf{Z} \sim U\{a, b\}$ denotes that $\mathbf{Z}$ is a uniformly distributed random variable between $a$ and $b$.

## II. SYSTEM, CHANNEL AND SIGNAL MODELS

In this section, the system, channel and signal models are presented for the multi-cell multi-user massive MIMO downlink with hybrid beamforming.

### A. System and channel models

The system model consists of $L$ co-channel cells, and each cell contains a massive MIMO base-station, which shares the same frequency and time resources to serve $K$ spatially distributed user nodes. The $l$th cell consists of $K$ single-antenna user nodes $(U_{lk})$ for $k \in \{1, \cdots, K\}$, and a base-station with $M$ antennas $(K \ll M)$. Moreover, each massive MIMO base-station is equipped with a hybrid beamformer in order to reduce the implementation complexity and power consumption in DSP. To this end, we consider $N$ transmit RF chains $K < N \ll M$ at the base-station. Therefore, we employ an analog phase-shifter at the base-station to reduce the burden on baseband precoder. Furthermore, the noise at the receivers is modeled as complex zero mean additive white Gaussian noise (AWGN).

The channel matrix between the base-station and $K$ user nodes in the $l$th cell is denoted by $\mathbf{F}_{ll}$ and is modeled as [1]

$$\mathbf{F}_{ll} = D_{ll}^{1/2} \hat{\mathbf{F}}_{ll},$$

where $\hat{\mathbf{F}}_{ll} \sim \mathcal{C}\mathcal{N}_{K \times M}(0_{K \times M}, \mathbf{I}_K \otimes \mathbf{I}_M)$ accounts for the independent, frequency-flat Rayleigh fading, and $D_{ll}$ is an $K \times K$ diagonal matrix, which captures the path-loss. The $k$th diagonal element of $D_{ll}$ can be denoted as $[D_{ll}]_{k,k} = \zeta_{ll,k}$ for $k \in \{1, \cdots, K\}$. Moreover, $\zeta_{ll,k}$ is assumed to be a constant over many channel coherence intervals and assumed to be known a priori as it changes very slowly with time. The channel matrix from base-station in the $l$th cell to the user nodes in the $i$th cell can be denoted as $\mathbf{F}_{li} = D_{li}^{1/2}\hat{\mathbf{F}}_{li}$, where $D_{li}$ captures the path-loss in corresponding channel.

### B. Acquisition of channel state information

In hybrid beamforming structure, a single RF chain is coupled to multiple antennas, and hence, each RF chain receives an inseparable sum of signals from its coupled antennas. In other words, the effective channel seen by the digital precoder is a product of the analog precoder and true channel matrix. Efficient and practically realizable channel estimation for hybrid beamforming is still an open problem. In this paper, we adopt a round-robin channel estimation technique proposed in [10]. During each round, $N$ base-station antennas are selected out of $M$ and trained by using $N$ RF chains. Since $M$ channel coefficients need to be estimated for each user, the aforementioned estimation process is repeated $[M/N]$ times in order to get all channels estimated. Though this method creates $M/N$ times more pilot overhead, it is easy to implement and provide tractable approach for full-dimensional channel estimation with only a few RF chains.

For the time division duplex (TDD) operation, the uplink channels are estimated at the base-station by using the uplink pilots transmitted by user nodes. To investigate the worst-case scenario, the same pilot sequence is assumed to be reused in all $L$ cells [1]. At the beginning of each coherence interval, user nodes simultaneously transmit the pilot sequence, and repeat the process for $[M/N]$ times until the base-station receives pilot signals from all $M$ antennas. The received pilot signal $\mathbf{Y}_p \in \mathbb{C}^M \times T_p$ at the base-station can be written as

$$\mathbf{Y}_p = \sum_{l=1}^{L} \sum_{k=1}^{K} \sqrt{P_p} \mathbf{f}_{ll,k}^T \Phi_{ll,k} + \mathbf{N}_p,$$

where $\mathbf{f}_{ll,k}$ is the uplink channel from the $k$th user in the $i$th cell to the $l$th cell base-station, and $\Phi_{ll,k} \in \mathbb{C}^{1 \times T_p}$ is the $k$th user’s pilot sequence, which satisfies the pairwise orthogonality property, i.e., $\Phi_{ll,k}^H \Phi_{ll,k}^H = 1$ and $\Phi_{ll,k}^H \Phi_{ll,k}^H = 0$ for $n \neq k$. Here, $P_p = T_p P_U$, where $T_p$ and $P_U$ are the portion of coherence interval utilized to transmit pilot signals.
and pilot transmit power, respectively. Moreover, in (2), \( N_p \) is an AWGN matrix with zero mean and unit variance complex Gaussian distributed entries. By first using the minimum mean square error (MMSE) estimation, and then invoking channel reciprocity, the relationship between the true downlink channel \( (f_{ll,k}) \) and its MMSE estimate \( (\hat{f}_{ll,k}) \) can be written as
\[
\hat{f}_{ll,k} = \sqrt{\theta_k} \hat{f}_{ll,k} + \sqrt{1 - \theta_k} E_{F_{ll,k}},
\]
where \( E_{F_{ll,k}} \) is the estimation error vector with zero mean and unit variance, and is independent of \( \hat{f}_{ll,k} \). Here, \( \theta_k = \zeta_{F_{ll,k}} / \left( \sum_{l=1}^L \zeta_{F_{ll,k}} + 1 / P_p \right) \) is the correlation coefficient between \( \hat{f}_{ll,k} \) and \( F_{ll,k} \).

In the case of frequency division duplex (FDD) mode, the downlink channels are estimated at the user nodes. Then, the quantized CSI is conveyed to the base-station by the user nodes by using a feedback channel. Specifically, depending on channel state, the \( k \)th user node selects an optimal codeword from a predetermined quantization codebook and feedbacks the index of the selected codeword. Thereby, the base-station recovers the quantized CSI from the same codebook. The relationship between the true channel and the quantized CSI can be approximated as follows [11]:
\[
\hat{f}_{ll,k} = \psi_k \hat{f}_{ll,k} + \sqrt{1 - \psi_k} E_{F_{ll,k}},
\]
where \( \hat{f}_{ll,k} = f_{ll,k} / |f_{ll,k}| \) and \( \hat{f}_{ll,k} = f_{ll,k} / |f_{ll,k}| \) are the true channel’s phase and quantized phase, respectively. Moreover, in (4), \( E_{F_{ll,k}} \) is the quantization error vector with uniformly distributed elements, and \( \psi_k = 1 - 2^{-\delta/(M-1)} \) is the associated correlation coefficient, where \( \delta \) channel feedback resolution.

It is worth noting that the estimated CSI in (3) and (4) for TDD and FDD, respectively, can be written in a unified form. However, this paper adapts TDD mode, and hence, \( F_{ll} \) can be alternatively written as follows [12], [13]:
\[
F_{ll} = \hat{F}_{ll} + E_{F_{ll}},
\]
where \( E_{F_{ll}} = [E_{F_{ll,1}}, \ldots, E_{F_{ll,K}}] \) and \( \hat{F}_{ll} = [\hat{f}_{ll,1}, \ldots, \hat{f}_{ll,K}] \). Here, the columns of \( \hat{F}_{ll} \) and \( E_{F_{ll}} \) are statistically independent and distributed as \( \mathcal{CN}(0, \hat{D}_{F_{ll}}) \) and \( \mathcal{CN}(0, \hat{D}_{F_{ll}} - \hat{D}_{F_{ll}}) \), respectively, where \( \hat{D}_{F_{ll}} \) is a diagonal matrix with the \( k \)th diagonal element is given by \( \zeta_{F_{ll,k}} = \theta_k \zeta_{F_{ll,k}} \) for \( k \in \{1, \ldots, K\} \).

C. Signal model

In this subsection, the downlink signal model for the \( ll \)th cell is presented. To begin with, the transmitted signal at the \( ll \)th cell base-station towards \( K \) user nodes can be written as
\[
x_l = \sqrt{\alpha_l} \hat{H}_l W_l x_i + n_K,
\]
where \( x_l = [x_{l1}, \ldots, x_{lk}, \ldots, x_{lK}]^T \) is the signal vector transmitted to \( K \) user nodes satisfying \( \mathbb{E}[x_l x_i^H] = I_K \), and \( P_l \) is the transmit power at the base-station. In (6), \( \hat{H}_l \in \mathbb{C}^{M \times N} \) and \( W_l \in \mathbb{C}^{N \times K} \) denote the analog beamformer and digital precoder, respectively, at the base-station in the \( ll \)th cell for \( l \in \{1, \ldots, L\} \). Moreover, \( \alpha_l \) is the power normalization factor and can be defined as
\[
\alpha_l = 1 / \sqrt{\mathbb{E}[\text{Tr}(\hat{H}_l W_l W_l^H H_l^H)]}.
\]

The received signal at the user nodes in the \( ll \)th cell can be written as
\[
y_l = \sum_{i=1}^L \sqrt{P_i} \alpha_i \hat{H}_l W_i x_i + n_l,
\]
where \( n_l \) is the AWGN vector at the user nodes satisfying \( \mathbb{E}[n_l n_l^H] = I_K \sigma_n^2 \).

D. Hybrid precoder design

In this subsection, the design of \( \hat{H}_l \) and \( \hat{W}_l \) is presented. In this design, it is permissible to vary both the amplitude and phase of the digital precoder \( \hat{W}_l \), whereas only the phase of the analog beamformer \( \hat{H}_l \) is allowed to change. Hence, the amplitude of each element in \( \hat{H}_l \) is constrained to \( 1/\sqrt{M} \). Moreover, the phase-only control of the analog beamformer is achieved by extracting the phase information of the conjugate transpose of the downlink channel, which is obtained via the uplink channel estimates at the base-station [2]. Due to the practical implementation challenges of continuous phase-shifters and to make the analysis more realistic, quantized phases are introduced. To this end, each of \( M \times K \) phases in \( F_{ll}^H \) is sampled by using \( \delta \) levels, each quantized to its nearest level based on the closest Euclidean distance. Therefore, the phases of the entries of \( \hat{H}_l \) can be assigned as
\[
\hat{\theta}_{m,k} = \begin{cases} \frac{2\pi\hat{n}/2^\delta}{\phi_{\text{rand}}} & k \leq K \\ \frac{2\pi m}{2^\delta} & K < k \leq N, \end{cases}
\]
where \( \delta \) is the number of quantization levels. Furthermore, \( \phi_{\text{rand}} \sim U[-\pi, \pi] \) and \( \hat{n} \) can be derived as
\[
\hat{n} = \arg\min_{n \in \{0, \pm 1, \ldots, 2\pi - 1\}} \left| \theta_{m,k} - \frac{2\pi m}{2^\delta} \right|,
\]
where \( \theta_{m,k} \) is the phase of \( \left[ F_{ll}^H \right]_{m,k} \) and is distributed as \( \theta_{m,k} \sim U[-\pi, \pi] \). Moreover, the error between the unquantized phase and quantized phase can be defined as
\[
\phi_{m,k} = \theta_{m,k} - \hat{\theta}_{m,k},
\]
where \( \phi_{m,k} \sim U[-\lambda, \lambda] \) and \( \lambda = \pi / 2^\delta \).

Next, the low-dimensional digital precoder \( \hat{W}_l \) is implemented by using the equivalent channel obtained from the product of analog beamformer and the estimated channel \( F_{ll} \) [2]. Thus, \( \hat{W}_l \) for ZF and MRT precoding can be derived as
\[
\hat{W}_l = \begin{cases} (\hat{F}_{ll}^H \hat{H}_l)^H (\hat{F}_{ll}^H \hat{H}_l) (\hat{F}_{ll}^H \hat{H}_l)^H \}^{-1} & \text{ZF} \\ (\hat{F}_{ll}^H \hat{H}_l)^H & \text{MRT}, \end{cases}
\]
where \( l \in \{1, \ldots, L\} \).

III. SUM RATE DEFINITIONS

In TDD massive MIMO systems, the channels are estimated only at the base-station by using the pilot sequences transmitted by the user nodes. Hence, user nodes are unaware of the corresponding CSI [1], [4], [14], and they rely on channel hardening for the downlink signal decoding. Therefore, only long-term channel statistics are assumed to be available at the user nodes. By extracting the \( l \)th row from (8), the received signal at the \( k \)th user node in the \( ll \)th cell can be rewritten as
The pilot sequence length is bounded as $T_p \geq K[M/N]$, symbol durations in length, via a genie-aided CSI acquisition technique. This assumption is made due to a lack of an efficient uplink channel estimation technique for the hybrid beamforming structures in the open literature. Moreover, this assumption is useful in deriving non-vanishing asymptotic sum rate expressions for theoretical performance comparisons.

### IV. Asymptotic Sum Rate Analysis

In this section, the asymptotic achievable sum rate expressions are derived for the hybrid beamforming with ZF and MRT precoders when the number of antennas at the base-station is allowed to grow without bound.

1) **Analog beamformer cascaded with the ZF precoder:**

To begin with, by substituting (12) into (16), and scaling the transmit power $^1$ at the base-station as $P_t = E_t / \sqrt{M}$, while keeping $E_t$ fixed, the achievable asymptotic sum rate for ZF precoding can be derived by letting $M \rightarrow \infty$ as (see Appendix A for the derivation)

$$
\tilde{R}_Z = \psi_1 \sum_{k=1}^{K} \log \left( 1 + \frac{P_1 \sigma_1^2 \mathbb{E} \left[ f_{l,k} \hat{H}_l \hat{w}_{li,k} x_i \right]^2}{\tilde{n}_k + 1} \right),
$$

where $\tilde{n}_k$ can be written as shown in (14) at the top of this page. In (13), the desired signal and the effective noise terms are uncorrelated. Thus, by using the fact that the uncorrelated Gaussian noise represents the worst-case scenario, the achievable sum rate can be tightly approximated as (15)

$$
\tilde{R}_l = \psi_1 \sum_{k=1}^{K} \log \left( 1 + \frac{P_1 \sigma_1^2 \mathbb{E} \left[ f_{l,k} \hat{H}_l \hat{w}_{li,k} x_i \right]^2}{\tilde{n}_k + 1} \right),
$$

where $\tilde{n}_k$ is the worst-case Gaussian approximation for the end-to-end signal-to-interference-plus-noise ratio at the $k$th user node in the $l$th cell. Moreover, in (15), $\psi_1$ is the pre-log factor, which captures the effective portion of coherence interval used for data transmission. By invoking the worst-case Gaussian approximation, $\tilde{n}_k$ can be derived as

$$
\tilde{n}_k = \sqrt{T_\text{Cx}} \left( \mathcal{E} \left[ f_{l,k} \hat{H}_l \hat{w}_{li,k} x_i \right] - \mathbb{E} \left[ f_{l,k} \hat{H}_l \hat{w}_{li,k} x_i \right] \right) + \frac{K}{q=1, q \neq k} \sum_{i=1, i \neq l}^{L} \sqrt{T_\text{Cx}} \alpha_i f_{l,k} \hat{H}_l \hat{w}_{li,q} x_i + \sum_{i=1, i \neq l}^{L} \sqrt{T_\text{Cx}} \alpha_i f_{l,k} \hat{H}_l \hat{w}_{li,q} x_i + n_k \text{ AWGN}
$$

where $\tilde{n}_k$ can be written as shown in (14) at the top of this page. In (13), the desired signal and the effective noise terms are uncorrelated. Thus, by using the fact that the uncorrelated Gaussian noise represents the worst-case scenario, the achievable sum rate can be tightly approximated as (15)

$$
\tilde{R}_l = \psi_1 \sum_{k=1}^{K} \log \left( 1 + \frac{P_1 \sigma_1^2 \mathbb{E} \left[ f_{l,k} \hat{H}_l \hat{w}_{li,k} x_i \right]^2}{\tilde{n}_k + 1} \right),
$$

where $\tilde{n}_k$ is the worst-case Gaussian approximation for the end-to-end signal-to-interference-plus-noise ratio at the $k$th user node in the $l$th cell. Moreover, in (15), $\psi_1$ is the pre-log factor, which captures the effective portion of coherence interval used for data transmission. By invoking the worst-case Gaussian approximation, $\tilde{n}_k$ can be derived as

$$
\tilde{n}_k = \sqrt{T_\text{Cx}} \left( \mathcal{E} \left[ f_{l,k} \hat{H}_l \hat{w}_{li,k} x_i \right] - \mathbb{E} \left[ f_{l,k} \hat{H}_l \hat{w}_{li,k} x_i \right] \right) + \frac{K}{q=1, q \neq k} \sum_{i=1, i \neq l}^{L} \sqrt{T_\text{Cx}} \alpha_i f_{l,k} \hat{H}_l \hat{w}_{li,q} x_i + \sum_{i=1, i \neq l}^{L} \sqrt{T_\text{Cx}} \alpha_i f_{l,k} \hat{H}_l \hat{w}_{li,q} x_i + n_k \text{ AWGN}
$$

**Remarks III.1:** The pre-log factor is defined as $\psi_1 = (T_c - T_p)/T_c$, where $T_c$ is the channel coherence interval, and $T_p$ is the pilot sequence length. According to the round-robin CSI acquisition technique discussed in Section II-B, the pilot sequence length is bounded as $T_p \geq K[M/N]$. Thus, this pre-log factor becomes asymptotically infinitesimal as the number of base-station antennas grows without bound. Consequently, under this round-robin CSI acquisition technique, the asymptotic achievable sum rate becomes infinitesimal due to the fact that $T_c$, which is the product of coherence time and coherence bandwidth, can be as few as two hundred symbol durations. However, in the sequel, it is assumed that the uplink channel can be estimated by using pilot sequences with $K \leq T_p \ll K[M/N]$ symbol durations in length, via a genie-aided CSI acquisition technique. This assumption is made due to a lack of an efficient uplink channel estimation technique for the hybrid beamforming structures in the open literature. Moreover, this assumption is useful in deriving non-vanishing asymptotic sum rate expressions for theoretical performance comparisons.

2) **Analog beamformer cascaded with the MRT precoder:**

By scaling the transmit power as $P_t = E_t / \sqrt{M}$, and by letting $M \rightarrow \infty$, the asymptotic achievable sum rate for MRT precoding can be derived as (see Appendix B for the derivation)

$$
P_M = \sum_{i=1, i \neq l}^{L} E_t E_l T_p \left[ \sqrt{\pi/4 \sin (\lambda)} \right]^2 \mathbb{E} \left[ f_{l,k} \hat{H}_l \hat{w}_{li,k} x_i \right]^2 / \mathbb{E} \left[ f_{l,k} \hat{H}_l \hat{w}_{li,k} x_i \right] + \sum_{i=1, i \neq l}^{L} \mathbb{E} \left[ f_{l,k} \hat{H}_l \hat{w}_{li,k} x_i \right] + n_k \text{ AWGN}
$$

where $\sin (\lambda) = \sin (\lambda) / \lambda$, $\lambda = \pi/2\delta$, and $\delta$ is the number of quantization levels.

1The pilot transmit power at the user nodes is also scaled as $P_U = E_t / \sqrt{M}$, where $E_t$ is fixed.
\[ \hat{R}_M^{\infty} = \psi_1 \sum_{k=1}^{K} \log \left( 1 + \frac{P_0^\infty}{\sigma_k^2} \right), \]  

\[ R_L^\infty = \frac{E_t E_i T_p}{\sqrt{\pi/4 \text{sinc}(\lambda)}} \zeta \sum_{k=1}^{K} \frac{\zeta_k}{\text{Tr}(D_{F,k}^2)}, \]  

where \( P_0^\infty \) and \( P_L^\infty \) can be written as

\[ P_0^\infty = E_t E_i T_p \left[ \sqrt{\pi/4} \text{sinc}(\lambda) \right]^2 \zeta_k \sum_{k=1}^{K} \frac{\zeta_k}{\text{Tr}(D_{F,k}^2)}. \]

\[ P_L^\infty = \sum_{i=1}^{L} E_t E_i T_p \left[ \sqrt{\pi/4} \text{sinc}(\lambda) \right]^2 \zeta_{i,k} \sum_{k=1}^{K} \frac{\zeta_k}{\text{Tr}(D_{F,k}^2)}. \]

3) Dimensional beamforming: For comparison purposes, the achievable asymptotic sum rates are derived for the full-dimensional digital precoding in which the analog beamformer is removed from the operation. To this end, by using the worst-case Gaussian approximation, the corresponding asymptotic achievable sum rate expressions for the full-dimensional ZF precoding and MRT can be derived as shown in (22a) and (22b), respectively, at the top of the previous page.

Remarks IV.1: As shown in (18) and (20), the achievable asymptotic sum rate performance is degraded due to the hybrid precoding implemented at the base-station. This observation can readily be verified by comparing (18) and (20) with the corresponding sum rates for the full-dimensional digital precoding in (22a) and (22b), respectively. To this end, the achievable asymptotic sum rate in (18) can be rewritten as in (23) at the top of the previous page. It is worth noting that the term \( \left[ \sqrt{\pi/4} \text{sinc}(\lambda) \right]^2 \) is less than unity for all possible values of \( \lambda \). Hence, by comparing (22a) and (23), it can be observed that the term \( \left[ \sqrt{\pi/4} \text{sinc}(\lambda) \right]^2 \) degrades the achievable sum rate of the hybrid beamforming over the full-dimensional digital precoding. Thus, the hybrid precoding provides a low complexity implementation at the base-station, while trading-off a portion of the achievable sum rate gains.

Furthermore, for strictly negligible noise power at the user nodes (i.e., \( \sigma_k^2 \rightarrow 0 \) or \( \sigma_k^2 \ll P_0^\infty \)), the achievable asymptotic sum rate expressions corresponding to both hybrid beamforming and full-dimensional digital precoding reduce to the sum rate expressions shown in (24) at the top of the previous page for MRT and ZF. Hence, when the noise powers at the user nodes are infinitesimal small compared to the residual interference due to pilot contamination, the detrimental effects of the quantized analog beamforming can be mitigated if \( N \gg K \). Consequently, the proposed hybrid precoding techniques can provide the same sum rates that are achieved by employing the full-dimensional digital precoding, while utilizing significantly smaller number of RF chains at the base-station.

V. Achievable Asymptotic Sum Rate Analysis - Perfect CSI

In this section, the achievable asymptotic sum rates corresponding to the genie-aided perfect CSI case are derived. To be more specific, both the analog beamformer and the digital precoder in (10) and (12), respectively, are designed by assuming the availability of perfect CSI. To begin with, by scaling the transmit power at the base-station as \( P_t = E_t / M \), while keeping \( E_t \) fixed, and by letting \( M \rightarrow \infty \), the asymptotic sum rates for the hybrid beamforming can be derived as

\[ R_2^\infty = \sum_{k=1}^{K} \log \left( 1 + \frac{E_t \pi \text{sinc}^2(\lambda)}{4 \text{Tr}(D_{F,k}^2) \sigma_k^2} \right), \]

\[ R_2^\infty = \sum_{k=1}^{K} \log \left( 1 + \frac{E_t \pi \text{sinc}^2(\lambda) \zeta_{F,k}}{4 \text{Tr}(D_{F,k}^2) \sigma_k^2} \right). \]

For the sake of completion, the asymptotic rates for the full-dimensional digital precoding with perfect CSI are derived as

\[ R_2^\infty = \sum_{k=1}^{K} \log \left( 1 + \frac{E_t \pi \text{sinc}^2(\lambda) \zeta_{F,k}}{4 \text{Tr}(D_{F,k}^2) \sigma_k^2} \right), \]

\[ R_2^\infty = \sum_{k=1}^{K} \log \left( 1 + \frac{E_t \pi \text{sinc}^2(\lambda) \zeta_{F,k}}{4 \text{Tr}(D_{F,k}^2) \sigma_k^2} \right). \]

Remarks VI.1: The asymptotic sum rate is degraded due to phase quantization of the analog beamforming even with the perfect channel knowledge at the base-station. By comparing (25a) and (25b) with respective to (26a) and (26b), it can be observed that the achievable asymptotic sum rate for the perfect CSI case is degraded by a factor of \( \pi \text{sinc}^2(\lambda)/4 \).

VI. Numerical Results

In this subsection, the simulation results are presented by assuming that the channel estimation can be performed in \( T_p = K \) symbol durations [10]. In Fig. 2, the average achievable sum rate is plotted against the number of antennas at the base-station. The asymptotic sum rate curves are plotted by using (18) and (20) for ZF and MRT, respectively, and compared against the Monte-Carlo simulations. Fig. 2 shows that the cumulative effects of phase quantization and pilot contamination significantly degrade the achievable sum rate for the proposed hybrid beamforming. These detrimental effects do not vanish even in the asymptotic antenna regime. For example, asymptotic sum rate losses of 0.8 bits/s/Hz and 1.0 bits/s/Hz are observed for hybrid precoding with ZF and MRT, respectively, when the number of cells increases from \( L = 2 \) to \( L = 4 \).
In this appendix, the derivation of the asymptotic achievable sum rate in (18) is outlined. To begin with, by using the law of large numbers, for mutually independent $n \times 1$ column vectors $\mathbf{a}$ and $\mathbf{b}$, whose elements are independent and identically distributed with zero mean and variances $\sigma_a^2$ and $\sigma_b^2$, respectively, it can be shown that [16]

$$\mathbf{a}^H/\sqrt{n} \xrightarrow{a.s.} \sigma_a^2$$

and

$$\mathbf{b}^H/\sqrt{n} \xrightarrow{a.s.} \sigma_b^2$$

as $n \to \infty$. However, by using the central limit theorem, it can be shown that

$$\mathbf{a}^H/\sqrt{\delta} \xrightarrow{d} \mathcal{CN}(0, \sigma_a^2)$$

and

$$\mathbf{b}^H/\sqrt{\delta} \xrightarrow{d} \mathcal{CN}(0, \sigma_b^2)$$

as $n \to \infty$, where $\delta$ denotes the convergence in distribution.

Next, by using the formation of $\tilde{\mathbf{H}}_{ii}$, the $(m, n)$th element of $\tilde{\mathbf{H}}_{ii}$ can be represented as

$$\frac{1}{\sqrt{M}} \exp(j\theta_{m,n})$$

where $\theta_{m,n}$ for $m \in \{1, \ldots, M\}$, $n \in \{1, \ldots, N\}$ is a discrete uniform random variable, i.e., $\theta_{m,n} \sim U(−\pi, \pi)$. Now, let $\tilde{h}_{i,p}$ and $\tilde{h}_{i,q}$ be the $p$th and the $q$th columns of $\tilde{\mathbf{H}}_{ii}$, respectively. Then,

$$\tilde{h}_{i,p}^H\tilde{h}_{i,q} = 1$$

as the opposite phases cancel out each other. Moreover, by using the central limit theorem, an asymptotic value of $\tilde{h}_{i,p}^H\tilde{h}_{i,q}$ for $p \neq q$ can be derived as

$$\lim_{M \to \infty} \tilde{h}_{i,p}^H\tilde{h}_{i,q} = \lim_{M \to \infty} \frac{1}{M} \sum_{m=1}^{M} \exp(j\theta_{m,p} - \theta_{m,q})$$

(29)

$$\xrightarrow{a.s.} \mathbb{E}[\exp(j\theta_{m,p})] \mathbb{E}[\exp(-j\theta_{m,q})] = 0.$$

Thus, as $M \to \infty$, it can be shown that $\tilde{h}_{i,p}^H\tilde{h}_{i,q} \xrightarrow{a.s.} \mathbf{I}_N$.

Then, the asymptotic values of $\tilde{\mathbf{F}}_{ii}$ and $\tilde{\mathbf{F}}_{ii}$ can be derived as follows: To this end, the magnitude of the $(k, m)$th element of $\tilde{\mathbf{F}}_{ii}$ can be alternatively written as

$$|\tilde{F}_{i,k,m}| = r_{k,m} \exp(-j\theta_{m,k})$$

(30)

where $r_{k,m} \sim R\left(R\left(\frac{\zeta_{F_i,k}}{2}\right)\right)$ and $\theta_{m,k}$ is given in (10). By treating $\tilde{F}_{i,p}$ and $\tilde{F}_{i,q}$ as the $p$th row and the $q$th column of $\tilde{\mathbf{F}}_{ii}$ and $\tilde{\mathbf{F}}_{ii}$, respectively, and by scaling the pilot transmit power as $P_U = E_U/\sqrt{M}$, an asymptotic value of $\tilde{F}_{i,p}\tilde{h}_{i,p}/M^{1/4}$ can be derived for $p \leq K$ as follows:

$$\tilde{F}_{i,p}\tilde{h}_{i,p}/M^{1/4} = \lim_{M \to \infty} \frac{1}{M^{3/4}} \sum_{m=1}^{M} r_{p,m} \exp(j\theta_{m,p} - j\theta_{m,p})$$

(31)

$$\xrightarrow{a.s.} \mathbb{E}[r_{p,m}] \mathbb{E}[\exp(-j\phi_{m,p})] / M^{3/4}$$

where the central limit theorem has been invoked in (a). In (31), $\lambda = \pi/2$ and $\phi_{m,p}$ is defined in (11). Moreover, by using an approach similar to that in (29), when $M \to \infty$, it can be shown that $\tilde{F}_{i,p}\tilde{h}_{i,q}/M^{1/4} \xrightarrow{a.s.} 0$ for $p \neq q$. By using similar steps to those used for deriving (31), an asymptotic expressions for $\tilde{F}_{i,p}\tilde{h}_{i,q}/M^{1/4}$ and $\tilde{F}_{i,p}\tilde{h}_{i,q}/M^{1/4}$ can be derived as follows:

$$\lim_{M \to \infty} \tilde{F}_{i,p}\tilde{h}_{i,q}/M^{1/4} \xrightarrow{a.s.} \sqrt{\pi} E_U T_p^2/4\zeta_{F_{i,p}} \exp(-j\phi_{m,p} / (4\sqrt{M})$$

(32a)

$$\lim_{M \to \infty} \tilde{F}_{i,p}\tilde{h}_{i,q}/M^{1/4} \xrightarrow{a.s.} 0.$$
When $M \to \infty$, $\hat{F}_{l,k} \tilde{H}_{l,k} / M^{1/4}$ and $\hat{F}_{l,k} \tilde{H}_{l,k} / M^{1/4}$ can be asymptotically approximated as

$$
\hat{F}_{l,k} \tilde{H}_{l,k} / M^{1/4} \xrightarrow{\alpha_l^2/\sqrt{M}} \frac{\sqrt{\pi E_U T_p}}{4 \text{sinc} (\lambda)} D_{F_l, I_{K,N}},
$$

(33a)

and

$$
\hat{F}_{l,k} \tilde{H}_{l,k} / M^{1/4} \xrightarrow{\alpha_l^2/\sqrt{M}} \frac{\sqrt{\pi E_U T_p}}{4 \text{sinc} (\lambda)} D_{F_l, I_{K,N}},
$$

(33b)

where $I_{K,N}$ is an $N \times N$ matrix whose first $K$ diagonal elements are unity, and the remaining elements are zero.

Next, we proceed with the derivation of the asymptotic achievable sum rate for the hybrid beamforming with ZF digital precoding. To begin with, by substituting (33a) into (12), $W_{Hl}$ can be simplified as

$$
M^{1/4} \tilde{W}_{Hl} \xrightarrow{\alpha^2_l/\sqrt{M}} \left[ \frac{\sqrt{\pi E_U T_p}}{4 \text{sinc} (\lambda)} \right]^{-1} \tilde{I}_{K,N} D_{F_l},
$$

(34)

By substituting (34) into (7), the term $\alpha^2_l / \sqrt{M}$ can be asymptotically approximated as

$$
\alpha^2_l / \sqrt{M} \xrightarrow{\alpha^2_l / \sqrt{M}} [ \frac{\sqrt{\pi E_U T_p}}{4 \text{sinc} (\lambda)} ] / 4 \text{Tr}(D_{F_l}).
$$

(35)

Then, by using (33b) and (34), $F_{l,k} \tilde{H}_{l,k} \tilde{W}_{Hl}$ can be asymptotically simplified as

$$
F_{l,k} \tilde{H}_{l,k} \tilde{W}_{Hl} \xrightarrow{\alpha^2_l / \sqrt{M}} D_{F_l} I_{K,N} D_{F_l}^{-1} \equiv I_{K,N},
$$

(36)

where step (a) is obtained by invoking the property $I_{K,N} D_{F_l} D_{F_l}^{-1} I_{K,N} = I_{K,N}$. By using (36), the terms $E \left[ f_{l,k} H_{l,k} \tilde{w}_{l,k} \right]$, $\text{Var} \left[ f_{l,k} H_{l,k} \tilde{w}_{l,k} \right]$ and $E \left[ \left| f_{l,k} H_{l,k} \tilde{w}_{l,k} \right|^2 \right]$ can be computed as follows:

$$
E \left[ f_{l,k} H_{l,k} \tilde{w}_{l,k} \right] \xrightarrow{\alpha^2_l / \sqrt{M}} \left[ E[I_{K,N}] \right]_{k,k} = 1,
$$

(37a)

$$
\text{Var} \left[ f_{l,k} H_{l,k} \tilde{w}_{l,k} \right] \xrightarrow{\alpha^2_l / \sqrt{M}} \left[ \text{Var}[I_{K,N}] \right]_{k,k} = 0,
$$

(37b)

$$
E \left[ \left| f_{l,k} H_{l,k} \tilde{w}_{l,k} \right|^2 \right] \xrightarrow{\alpha^2_l / \sqrt{M}} \left[ E[I_{K,N}] \right]_{k,k} = 0.
$$

(37c)

Moreover, $E \left[ \left| f_{l,k} H_{l,k} \tilde{w}_{l,k} \right|^2 \right]$ can be simplified as shown in (38) at the top of this page. By substituting (37a)-(37c) and (38) into (17), and by scaling the power as $P_l = E_l / \sqrt{M}$ and $P_U = E_U / \sqrt{M}$, the power terms in (16) can be rewritten as

$$
P_l \alpha^2_l \left[ E \left[ f_{l,k} H_{l,k} \tilde{w}_{l,k} \right] \right] \xrightarrow{\alpha^2_l / \sqrt{M}} \lim_{M \to \infty} E_l \left( \alpha^2_l / \sqrt{M} \right),
$$

(39a)

$$
P_U \alpha^2_l \left[ E \left[ f_{l,k} H_{l,k} \tilde{w}_{l,k} \right] \right] \xrightarrow{\alpha^2_l / \sqrt{M}} \lim_{M \to \infty} E_U \left( \alpha^2_l / \sqrt{M} \right),
$$

(39b)

where the terms $P_l \alpha^2_l \text{Var} \left[ f_{l,k} H_{l,k} \tilde{w}_{l,k} \right]$ and $E \left[ P_l \right]$ asymptotically converge to zero. By substituting (35) and by evaluating the limits in (39a) and (39b), the asymptotic values of the power terms can be derived as shown in (19). Thereby, the asymptotic achievable sum rate can be derived as in (18).

**APPENDIX B**

**DERIVATION OF SUM RATE IN (20)**

In this appendix, the derivation of asymptotic achievable sum rate for MRT in (20) is outlined. By substituting (33a) into (12) and (7), $W_{Hl} / M^{1/4}$ and $\alpha^2_l / \sqrt{M}$ can be asymptotically approximated as

$$
W_{Hl} / M^{1/4} \xrightarrow{\alpha^2_l / \sqrt{M}} \left[ \frac{\sqrt{\pi E_U T_p}}{4 \text{sinc} (\lambda)} \right]^{1/2} I_{K,N} D_{F_l},
$$

(40)

$$
\alpha^2_l / \sqrt{M} \xrightarrow{\alpha^2_l / \sqrt{M}} \left[ \frac{\sqrt{\pi E_U T_p}}{4 \text{sinc} (\lambda)} \right] \text{Tr}(D_{F_l}) / 4^{-1}.
$$

(41)

By using (33b) and (40), and following steps similar to those in (36), $F_{l,k} H_{l,k} W_{Hl}$ can be asymptotically simplified as

$$
F_{l,k} H_{l,k} W_{Hl} \xrightarrow{\alpha^2_l / \sqrt{M}} A D_{F_l}^2,
$$

(42)

where $A = E_U T_p / \sqrt{\pi} \text{sinc}^2 (\lambda) / \sqrt{M}$. Then, by using (42),

$$
E \left[ f_{l,k} H_{l,k} \tilde{w}_{l,k} \right], \text{Var} \left[ f_{l,k} H_{l,k} \tilde{w}_{l,k} \right], E \left[ \left| f_{l,k} H_{l,k} \tilde{w}_{l,k} \right|^2 \right]
$$

and

$$
E \left[ \left| f_{l,k} H_{l,k} \tilde{w}_{l,k} \right|^2 \right] \xrightarrow{\alpha^2_l / \sqrt{M}} A \text{Tr}(D_{F_l}) / 4^{-1}.
$$

(43a)

(43b)

(43c)

(43d)

Then, by substituting (43a)-(43d) into (16) and by scaling the transmit power as $P_l = E_l / \sqrt{M}$ and $P_U = E_U / \sqrt{M}$, the asymptotic power terms for MRT can be computed as shown in (21). Thereby, the desired achievable sum rate can be evaluated as given in (20).