IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 9, 2022, accepted January 31, 2022, date of publication February 7, 2022, date of current version February 14, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3149301

Precoding and Beamforming Techniques in mmWave-Massive MIMO: Performance Assessment

TEWELGN KEBEDE[1], YIHENEW WONDIE[1], JOHANNES STEINBRUNN[2], HAILU BELAY KASSA[3], AND KEVIN T. KORNEGAY[3]

[1]Addis Ababa Institute of Technology, Addis Ababa University, Addis Ababa 3614, Ethiopia
[2]Faculty of Electrical Engineering, Kempten University of Applied Science, 87435 Kempten, Germany
[3]Department of Electrical and Computer Engineering, Morgan State University, Baltimore, MD 21251, USA

Corresponding author: Tewelgn Kebede (tewelgn@gmail.com)

ABSTRACT Massive MIMO and mmWave communication are the technologies for achieving 5G design goals. Fortunately, these two technologies share a symbiotic integration. As a result, amalgamating mmWave communications with massive MIMO forms,mmWave-massive MIMO," which significantly improves spectral and energy efficiency. It also achieves high multiplexing gains and increases mobile network capacity. However, massive MIMO, mmWave communications, and mmWave-massive MIMO systems have been studied independently. Consequently, this article explores the ideas, performances, comparisons, and discussions of these three 5G technologies jointly, considering their precoding and beamforming methods. On the other hand, the complexity of these technologies increases when a large number of antennas and radio frequencies (RFs) are used. Thus, several investigations are going on to search for the appropriate precoding and beamforming strategies with low cost, power, and complexity. Therefore, massive MIMO linear precoding techniques such as zero-forcing, maximum ratio transmission, regularized zero-forcing, truncated polynomial expansion and phased zero forcing are addressed in this work. In addition, the most common non-linear precoding schemes: dirty-paper coding, Tomlinson-Harashima, and vector perturbation, are presented. Furthermore, a detailed discussion of the beamforming techniques called analog, digital, and hybrid analog-digital beamforming schemes is included. We also examine the potential of hybrid analog-digital beamforming with its fully-connected and sub-connected architecture approaches in making mmWave massive MIMO a reality. We evaluate their performance with the parameters: bit error rate, signal to noise ratio, complexity, spectral efficiency, and energy efficiency. According to the analytical and simulation results, the partially-connected hybrid analog/digital beamforming architecture offers better all-over performance for mmWave-Massive MIMO communications by compromising: power consumption, cost, complexity, and performance. Finally, the potential future directions in mmWave-massive MIMO precoding and beamforming challenges are addressed.

INDEX TERMS Beamforming, hybrid, massive MIMO, mmWave, mmWave-massive MIMO, precoding.

I. INTRODUCTION

Wireless technology networks have become increasingly prevalent in many aspects of everyday life, with alarming growth over the last several decades. Because of the proliferation of new applications and services, users' demand have increased. The 5^{th} generation (5G) wireless network emerges at this moment, committed to opening the screen with a wide range of informational durations and providing excellent customer satisfaction. As a result, investigations on key 5G technologies are being conducted in telecommunication networks for the corresponding verification and validation [1].

Massive MIMO realizes 5G goals using a large number of antennas. The physical sizes of the antennas are small and inexpensive, which improve both bandwidth and power consumption [2]. Deploying a large number of antennas at the basestaion (BS) helps to direct the transmission energy into

The associate editor coordinating the review of this manuscript and approving it for publication was Miguel López-Benítez.

a smaller region of space.With multiple array antennas, objects that were random before tend to be deterministic, which averages out the influence of small-scale fading. Moreover, as the number of BS antennas increases, the random channel vectors between users become orthogonal pair-wise.

Millimeter wave communication is also another technology to realize 5G goals. Due to the increased number of users and demand for high data rates, the classic 6 GHz frequency band is depleted [3]. Thereby, the mmWave spectrum, which has a frequency range of 30 - 300 GHz and a wavelength of about 10mm - 1mm, is expected to meet these demands (Fig. 1).

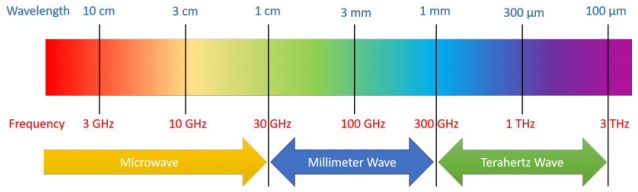

FIGURE 1. Frequency ranges.

Millimeter-wave communication has a substantial quantity of unutilized bandwidth, which has a capacity to connect millions of devices simultaneously. On the other hand, the mmWave band's high path-loss makes long-distance wireless communication difficult. Massive MIMO antennas are expected to be deployed in mmWave communication systems to compensate for path loss and extend coverage [4]. Luckily, massive MIMO and mmWave communications have symbiotic relationships in many ways [5]. The short wavelength of mmWave frequencies makes them compelling for massive MIMO. Because the physical size of the antenna arrays can be significantly reduced, making smaller cell sizes are promising for mmWave short-range communications. Furthermore, massive MIMO's large antenna gains also effectively alleviate mmWave signal severe path loss. So, integrating these two technologies enables the target to achieve a 1000-fold expansion in 5G efficiency. That is, by merging mmWave and massive MIMO, we get ''mmWave-massive MIMO'' that enables us to achieve the benefits of both significant area coverage and localized small cell hot-spots.

With the many number of BS antennas and simple precoding at the basestation, uncorrelated noise and intra-cell interference disappear entirely [6]. Massive MIMO may include precoding at the transmitter and decoding at the receiver, subject to indisputable requirements. Precoding consists of two parts: choosing the directivity (beamforming) and refering the transmit power (power allocation) [106].Thus, this article applies beamforming for referencing directivity and precoding for transmit-power allocation that achieves high capacity through spatial multiplexing of many users.

Beamforming strategies handles large antenna arrays precisely during transmission and reception, because mmWave frequencies are more directional than lower frequencies [3]. The transmit beamforming and receiving combiner vectors are designed based on channel state information (CSI). As a result, the layout of an effective beamforming method is obtained through optimal channel estimation. The primary categories of beamforming architectures are analog, digital, and hybrid beamforming architectures (Figure 3) [39], [40]. In traditional multi-user MIMO systems, fully-digital precoding is the typical approach to adjust the amplitudes and phases of the transmitted signals. The digital precoding approaches are further classified as linear and non-linear [7], [8], [41]. Among the linear precoding schemes that can be used to manage interference between single antenna users are: zero-forcing (ZF), maximum ratio transmission (MRT), regularized zero-forcing (RZF), truncated polynomial expansion (TPE), and phased zero-forcing (PZF) [8]–[11]. On the other hand, non-linear precoding approaches are: dirty-paper coding (DPC), Tomlinson-Harashima (TH), and vector perturbation (VP) [12]. As far as directivity is concerned, digital beamforming is further classified as fully-connected and partially-connected hybrid architectures. In a fully-connected structure, each RF chain is connected to all antenna elements, while in a partially-connected architecture, each RF chain only connects to a subset of antennas [103].

Downlink precoding is used with perfect CSI to optimize link performance. In addition, by outfitting the basestation with a large number of antennas, linear precoding techniques operate nearly the same as non-linear precoding methods [12]. Block diagonalization (BD) is used as a generalization of ZF for users with multiple receiving antennas [13]. Block diagonalization estimates a user's precoding matrix so that the precoding matrix's subspace lies in the null space of all other users, which eliminates inter-user interference [105]. However, in the fully-digital structure, the number of radio-frequency (RF) chains is equal to the number of antennas, which results in prohibitive hardware complexity and cost for massive MIMO systems. Analog-only precoding schemes were proposed in [14] to address complexity and cost issues. In analog structures, only lower-cost analog phase shifters are utilized to control the phases of the transmitted signals. However, phase shifters impose a constant modulus constraint on the precoding matrix entries, leading to fewer degrees of freedom and poorer precoding performance than fully-digital (FD) precoding. Another approach is selecting a subset of transmit antennas using simple switches, which does not provide full spatial diversity [15].

Several study papers have been published to discuss various issues related to massive MIMO systems [16]–[18]. In [9], a comprehensive survey of linear precoding techniques for massive MIMO systems with different cell approaches is introduced. It also handles some design constraints and practical precoding implementations. A survey of the mmWave massive MIMO system challenges and benefits is introduced in [19]. The authors in [20] conducted a comprehensive survey of linear precoding techniques for massive MIMO systems under a single-cell (SC) scenario. The performance of various linear precoding techniques is compared and analyzed in terms of sum rate and spectral efficiency. In [21],

a comprehensive review of various prominent mmWave massive MIMO systems, like multiple access technologies, hybrid precoding and combining, cell-free massive MIMO, and non-orthogonal multiple access technologies, are presented. In [22], a survey of the different analog, digital and hybrid analog/digital beamforming schemes are described by deploying CSI. The hybrid beamforming encounters the feasible limits of the RF chains. Besides, the hybrid beamforming design is considered as a trade-off between performance and complexity in the various application designs and channel characteristics. The authors in [23], consider the design of fully-connected and partially-connected hybrid architectures between the optimal fully-digital and the hybrid system. On the other hand, the survey in [24] only focuses on linear precoding techniques in massive MIMO systems for different cell scenarios.

The articles mentioned above discuss massive MIMO, mmWave communications, and mmWave massive MIMO systems independently for their respective beamforming and precoding systems. Most of these researchers focus only on parts of linear or non-linear and/or beamforming schemes. None of them jointly reviews precoding and/or beamforming techniques with their respective technologies together.

A. PAPER CONTRIBUTION

In this performance evaluation, we examine a comprehensive assessment of massive MIMO, mmWave Communication, and mmWave massive MIMO systems in terms of beamforming and precoding techniques. Performance and complexity trade-offs and the practical implementation of analog, digital, and hybrid beamforming architectures, targeting on 5G wireless networks, are performed. The digital precoding techniques, linear and non-linear, as well as their sub-categories, are also investigated. The performance of digital and hybrid analog-digital beamforming architectures is also covered, with fully-connected and partially-connected schemes.

Therefore, to the best of the authors' knowledge, there has been no performance evaluation available in the scientific world that has evaluated 5G technologies known as massive MIMO, mmWave communications, and mmWave-Massive MIMO technology jointly with respect to their precoding and beamforming schemes at this level of detail and wide coverage areas (linear precoding, non-linear precoding, analog, digital, and hybrid beamforming). Of course, there are a plethora of surveys available, such as [24], [25], and [9], that investigate these three technologies separately. Consequently, to achieve 5G goals, these three technologies are addressed together in terms of precoding and beamforming schemes. Besides, the emerging trends, challenges, and proposed approaches for mmWave-massive MIMO systems in the 5G era are investigated.

Thus, the contributions of this work are summarized as follows.

1) Presentation of a detailed analysis, comparison, merits, and demerits of the existing linear precoding techniques called: zero-forcing, maximum ratio transmission, regularized zero-forcing, truncated polynomial expansion and phased zero-forcing for downlink transmission.
2) Formulation of the analytical expressions of the performance metrics: bit error rate, spectral efficiency, and energy efficiency to evaluate precoding and beamforming techniques.
3) Review of massive MIMO non-linear precoding techniques such as dirty-paper coding, Tomlinson-Harashima, and vector perturbation with their performance comparisons and potential solutions.
4) Survey the analog, digital, and hybrid analog/digital beamforming for both mmWave communications and mmWave-Massive MIMO systems.
5) Review of the fully-connected and partially-connected architectures of hybrid analog-digital beamforming.
6) Finally, this work evaluates a comprehensive overview, comparison, discussion, and analyses of the performance of linear precoding, non-linear precoding, analog, digital and hybrid beamforming. Besides, it presents current trends with the possible research directions of both precoding and beamforming architectures for the three 5G wireless enabling technologies of massive MIMO, mmWave, and mmWave-massive MIMO communications.

B. PAPER ORGANIZATION

The rest of the paper is organized as follows. Massive MIMO is described in Section II. MmWave communications and mmWave-Massive MIMO communications with their precoding, beamforming, system, and channel models are described in Sections III and IV, respectively. Finally, Section V concludes the paper following challenges and open issues in section VI.

II. MASSIVE MIMO COMMUNICATION TECHNOLOGY

Massive MIMO refers to the use of antenna arrays with a large number of antenna elements for large-scale high-gain adaptive beamforming and multi-user spatial multiplexing applications [26]. It has been studied for conventional sub-6 GHz systems and is crucial for mmWave systems that require high directivity. Moreover, higher frequency bands offered by mmWave systems grant the ability to design antenna arrays with a considerable number of antennas [27]. That is, large-scale antenna arrays are system requirements for 5G higher frequency band communication systems such as mmWaves to tackle the weak propagation characteristics of these bands [28]. Therefore, the prospective benefits of massive MIMO are capacity and link reliability, high spectral and energy efficiency, security enhancement and robustness improvement, high degree of freedom, cost efficiency, and simple signal processing [29]–[31]. Similarly, massive MIMO also has a few key features: linear signal processing, favorable propagation, channel hardening, small physical size antennas with low power consumption, and scalability [16], [32]–[34].

To make this technology reach the efficient implementation stage, the challenges that need further investigation are channel estimation, pilot contamination, antenna array design, hardware impairment, precoding, and detection [35], [36].

A single-cell downlink massive MIMO system model is used to analyze massive MIMO performance (Figure 2). The BS has M antennas to support K single-antenna users that share the same time-frequency resource.

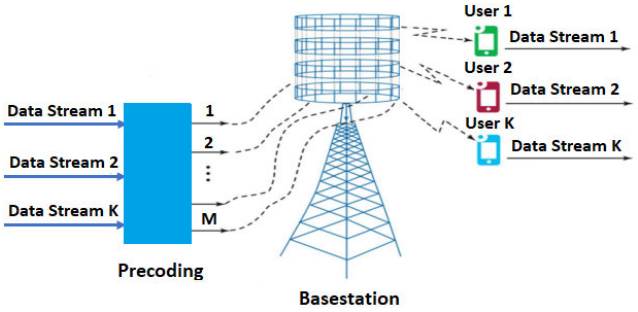

FIGURE 2. Downlink massive mimo system model.

If x represents the complex valued $M \times 1$ transmitted signal vector from M antennas, then the $K \times 1$ received signal vector y at the users is [8], [109]

$$y = \sqrt{p_d} H x + n \tag{1}$$

where $H \in \mathbb{C}^{K \times M}$ represents a channel matrix between M basestation antennas and K users. At the basestation, the channel is supposed to be Rayleigh fading and ergodic, with perfect channel state information. The elements of H are considered as independent and identically distributed (i.i.d) complex Gaussian random variables with zero mean and unit variance. The signal received by the k^{th} user after using the linear precoding is $x = Ws$ is the precoded signal at the basestation, $s \in \mathbb{C}^{K \times 1}$ is the information bearing signal with $\mathbb{E}\{ss^H\} = I_K$, $W \in \mathbb{C}^{M \times K}$ is the precoding matrix at the BS and p_d is the downlink transmit power for user k which is fixed. The equal power allocation policy tends to be optimal as the number of BS antennas grows without limit. That means large antenna arrays produce the same effect as increasing the transmit power, which cuts down the transmit power for uplink and downlink [107]. n is additive white Gaussian noise vector at the users. Then, the received signal by the k^{th} user after using the linear precoding is [10].

$$y_k = \sqrt{p_d} h_k w_k s_k + \sum_{i=1, i \neq k}^{K} \sqrt{p_d} h_k w_i s_i + n_k \tag{2}$$

where $\sqrt{p_d} h_k w_k s_k$ is the desired signal for user k and $\sum_{i=1, i \neq k}^{K} \sqrt{p_d} h_k w_i s_i$ is the multi-user interference signal. The signal to interference plus noise ratio (SINR) of user k is given by [10]

$$\text{SINR}_k = \frac{p_d |h_k w_k|^2}{p_d \sum_{i=1, i \neq k}^{K} |h_k w_i|^2 + 1} \tag{3}$$

which is a function of the channel h_k and transmit precoding vector w_k. If the interference cancellation is not made correctly, the receiver does not obtain the desired output [37]. Different precoding techniques have been proposed to mitigate inter-user interference, which is classified as linear and nonlinear precoding techniques (Figure 3) [8].

A. PERFORMANCE METRICS OF MASSIVE MIMO LINEAR PRECODING TECHNIQUES

The metrics to measure the performance of wireless technology are: bit error rate, signal to noise ratio, achievable sum rate, and energy efficiency [52].

1) BIT ERROR RATE OF LINEAR PRECODING

The bit error rate (BER) quantifies the number of errors in received bits over a transmission channel. This occurs when bits are altered due to interference, noise, distortion, or synchronization errors. It is also expressed as the ratio of the number of bits in error to the total number of bits transmitted over a given time interval [51].

The BER of a download single-cell massive MIMO communication system is determined by the modulation scheme used for transmission. The average BER for the k^{th} user in a massive MIMO system using ZF precoding and gray-coded square QAM modulation is described as [53]–[55]

$$P_e(\gamma_k) \approx \frac{c_N}{2d_N} \frac{\Gamma(\tau + \frac{1}{2}) / \Gamma(\tau + 1)}{(\gamma_k d_N^2 + 1)^{(\tau + \frac{1}{2})} \sqrt{\pi \gamma_k}} \tag{4}$$

where $\tau = M - K$ is the degree of freedom, $\gamma_k = \frac{P_T}{K\sigma^2}$ is the transmission SNR of the user, P_T is the total transmission power at the BS which is divided equally for each user and $\Gamma(.)$ is the Gamma function. The constant c_N and d_N is derived from the modulation level as

$$(c_N, d_N) = \begin{cases} (1, 1) & N = 2 \\ (2\frac{1 - 1/\sqrt{N}}{\log_2(\sqrt{N})}, \sqrt{\frac{3/2}{N-1}}) & N \geq 4. \end{cases} \tag{5}$$

The effect of $\Gamma(\tau + \frac{1}{2}) / \Gamma(\tau + 1)$ is negligible and if we omit it, the BER is approximated as

$$P_e(\gamma_k) \approx (\gamma_k d_N^2 + 1)^{-(\tau + \frac{1}{2})} \gamma_k^{-\frac{1}{2}}. \tag{6}$$

The result in (6) describes how deploying a large number of antennas at the BS improves the BER but makes it worse as the number of scheduled users increases [56].

2) ACHIEVABLE SUM RATE OF LINEAR PRECODING

Another metric used to analyze the performance of massive MIMO systems is the achievable rate. It specifies the lowest spectral efficiency that a MIMO system can achieve over a fading channel. In this regard, the Shannon channel capacity theory governs the achievable rate per user in a single-cell downlink massive MIMO network which is expressed as [57].

$$R_k = B \log_2(1 + \text{SINR}_k) \tag{7}$$

and the achievable sum rate of the system with K users is [10], [58]

$$R_s = B\sum_{k=1}^{K}\log_2(1+\text{SINR}_k) \tag{8}$$

where B denotes the bandwidth of the system.

3) POWER CONSUMPTION IN MASSIVE MIMO

The amount of power transmitted and the circuit's power consumption determine the total power consumption of the downlink massive MIMO system [2]

$$P_t = \mu\sum_{k=1}^{K}p_k + M P_c + P_{\text{fix}} \tag{9}$$

where μ is the inverse of the power amplifier efficiency at the basestation, p_k is downlink transmitter power allocated to each user, P_c is the constant circuit power consumption per antenna which contains power dissipation in the transmit filter, mixer, frequency synthesizer, and digital-to-analog converter, P_{fix} is the fixed power consumption at the basestation which is independent of the number of transmit antennas.

4) ENERGY EFFICIENCY OF LINEAR PRECODING

Energy efficiency is the ratio in which the achievable rate is compared with the system's associated power consumption (bits/joule) [59], [60].

$$\text{EE} = \frac{\text{Achievable sum rate}}{\text{Total power consumption}}. \tag{10}$$

Inserting (8) and (9) into (10), the energy efficiency of a downlink massive MIMO system is given by

$$\text{EE} = \frac{B\sum_{k=1}^{K}\log_2(1+\text{SINR}_k)}{\mu\sum_{k=1}^{K}p_k + M P_c + P_{\text{fix}}}. \tag{11}$$

Typically, as transmit power increases, so does the sum rate. However, there is a tradeoff between transmit power and energy efficiency, and thus it increases until some transmit power is reached and then decreases.

B. LINEAR PRECODING TECHNIQUES

Massive MIMO systems have sparked a great deal of research interest in mobile communications. Among massive MIMO evaluation instances and design challenges, employing precoding techniques at the basestation reduces signal processing complexity [9]. Precoding techniques reduce interference and confront signal processing complexity [12]. Precoding entails sending each data signal from all antennas but with varying amplitude and phase to direct the signal spatially. The purpose of precoding is to take advantage of the CSI available at the transmitter. Its concern is the incapability to mitigate IUI [38]. To reduce IUI, several precoding approaches have been applied. There are primarily three categories of beamforming architectures that are commonly explored [39], [40]: analog, digital, and hybrid beamforming architecture as shown in Figure 3.

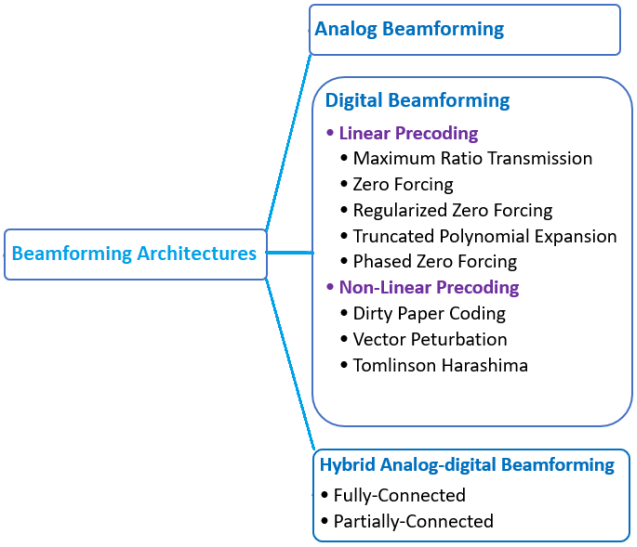

FIGURE 3. Beamforming techniques in massive MIMO systems [8], [9], [12], [37], [38].

The digital precoding methods are further divided into linear and non-linear schemes [8], [41]. In linear precoding, users are assigned different precoding matrices at the transmitter. Each antenna's transmit signal is generated separately in the digital baseband, which gives full flexibility in the signal generation. In linear precoding, as the name implies, data is transmitted linearly, which makes to have lower complexity and good performance.

Non-linear precoding techniques, on the other hand, are more complicated to implement but have a more significant potential than linear schemes [37]. According to the theoretical results, non-linear techniques are optimal, allowing for the maximum system sum rates. However, they are more expensive and complex, which makes them impractical in real-world systems [38]. If the goal is to achieve low complexity, then non-linear techniques should be replaced in favor of linear ones. As demonstrated in [37], linear precoders perform admirably when the number of BS antennas grows large.

The massive MIMO linear precoding techniques' precoding matrix (W) is summarized below [8]–[10], [23], [42]:

$$W = \begin{cases} H^H, & \text{MRT} \\ H^H(HH^H)^{-1}, & \text{ZF} \\ \beta(HH^H+\alpha I_K)^{-1}H^H, & \text{RZF} \\ \sum_{j=0}^{j-1} w_j(H^T H^*)^j H^T, & \text{TPE} \\ H_{eq}^H(H_{eq}H_{eq}^H)^{-1}\lambda, & \text{PZF} \end{cases}$$

1) ZERO FORCING PRECODING TECHNIQUE

Zero forcing (ZF) is a spatial signal processing technique. It has a large number of transmitting antennas that cancel out multi-user interference signals. In other words, it is used to

average out inter-user interference at each user. Compared to other linear precoding types, it is also known as full-complex zero forcing. Its precoding matrix is intended to minimize all interference among the user streams, making an optimal linear precoding scheme in massive MIMO systems [11]. ZF is suitable for low-noise or high-power implementations. As far as the channel is ill-conditioned, resulting in a low SNR at the receivers, the power of the precoded signal is substantially improved. IUI is removed at the cost of increased transmit power citeruim43. Interference is eliminated by sending signals in the ''direction'' of the projected user with nulls in the ''direction'' of other users, as cited by [44]. As a result, ZF is also known as null-steering. Its precoding matrix is generated by employing the inverse of the channel matrix as a transmission filter and applying the pseudo-inverse of the channel matrix. Channel inversion is generally preferred due to its low complexity and lack of sensitivity to channel estimation errors (Figure 4). On the other hand, channel inversion has a polynomial complexity in terms of both the number of concurrent users served and the number of antennas at the basestation.

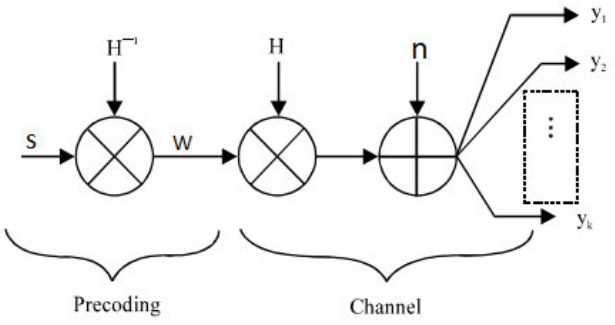

FIGURE 4. Massive MIMO system with channel inversion [45].

The IUI cancellation is conducted with this type of precoding by pre multiplying the signal targeted directly at the terminal by a precoding matrix. As a result, the ZF precoding matrix used by the BS is [10]

$$W_{\mathrm{ZF}} = H^{\mathrm{H}}(HH^{\mathrm{H}})^{-1}. \tag{12}$$

It's the pseudo-inverse of the channel matrix H. In the absence of additive noise, ZF precoding is an optimal precoding strategy [9]. If additive noise is present, this precoding technique amplifies the noise effect.

2) MAXIMUM RATIO TRANSMISSION PRECODING TECHNIQUE

MRT is also known as matched filter. Its goal is to optimize the SINR at the expected target by using the conjugate transpose of the channel matrix. As a result, the MRT precoding matrix used by the *BS* is therefore expressed as [8]

$$W_{\mathrm{MRT}} = H^{\mathrm{H}} \tag{13}$$

MRT precoder is preferred because of the absence of channel matrix inversion. In addition, it is the simplest linear precoder. Since MRT computes its weight by taking the complex conjugate of a signal, it only requires a bit-flip and an adder.

Given an appropriate SNR requirement, ZF prefers MRT as the number of antennas at the massive array M increases, making MRT an ideal precoder. Unlike ZF, MRT does not attempt to cancel IUI but boosts the symbol's power.

When using MRT instead of ZF precoders, the polynomial complexity of ZF-based beamforming in multi-user systems with a large number of antennas is overwhelming [38]. However, as demonstrated in [46], the optimality of conjugate beamforming for ZF should not be assumed in actual massive MIMO systems, as it would occur only in specific scenarios. MRT precoding, for example, outperforms only ZF when the number of concurrent users is very high, users have high mobility, or the hardware capacity at the basestation is very low. Therefore, the MRT precoder is preferred because of the absence of channel matrix inversion. Since MRT computes its weight by taking the complex conjugate of a signal, it only requires a bit-flip and an adder.

3) REGULARIZED ZERO FORCING PRECODING TECHNIQUE

Instead of a direct channel reversal, RZF precoding employs regularized channel inversion [43]. The regularized channel inversion is the function of balancing the signal-magnitude alignment and noise suppression which is a method of tuning the function by adding an additional penalty term in the RZF expression, α as shown in expression (14). Channel inversion is usually favored because it is less prone to channel estimation errors. Inter-user interference is reduced at the cost of transmit power enhancement. As a result, if channel inversion precoding is used to serve the K homogeneous users, transmission along the eigen-vector associated with the ill-behaving eigen-value would consume a significant amount of power. As a result, sum-rate performance does not scale linearly with the number of users [47].

The RZF precoder is typically obtained by minimizing the mean square error (MSE) among both transmitted and received symbols, also known as the minimum MSE (MMSE) [9]. It has been regarded as the state-of-the-art linear precoder for MIMO wireless communications systems due to its ability to trade off the advantages of both MRT and ZF precoders [48].

Thus, according to [43] and [44], the RZF precoding is

$$W_{RZF} = \beta(HH^{H} + \alpha I_K)^{-1}H^{H} \tag{14}$$

where β is the power normalization parameter and $\alpha = K/p_d$ is the regularization factor to be optimized. The implementation of regularization factor α also provides an extra degree of freedom for beamforming optimization [43]. At high SNR_s (α small), the *RZF* pre-coder tends to the ZF pre-coder, while at low SNR_s (α large), the MMSE pre-coder approaches to the MRT pre-coder [44]. I_K is KxK identity matrix, and p_d is downlink transmitter power assigned to each user.

The RZF precoding matrix is obtained by inverting a regularized channel matrix. However, as the number of antennas increases, so does the system's complexity.

4) TRUNCATED POLYNOMIAL EXPANSION PRECODING

TPE-based precoding has recently been presented to reduce the computational complexity of the RZF precoder while keeping comparable performance [9], [49]. It uses a polynomial expansion with J terms instead of RZF precoder matrix inversion, which is then truncated. TPE is thus obtained by approximating the inverse matrix in RZF with a $(J-1)$ -degree matrix polynomial, allowing for low complexity multi-stage hardware implementation. When J is changed, there is a performance transition with MRT ($J = 1$) and RZF ($J = min(M, K)$). Because the hardware complexity of TPE precoding is proportional to J, the hardware complexity is tailored to the deployment scenario [50]. The RZF precoder is typically obtained by minimizing the mean square error (MSE) among both transmitted and received symbols, also known as the minimum MSE (MMSE) [9]. The RZF precoder has been regarded as the state-of-the-art linear precoder for MIMO wireless communications systems due to its ability to trade off the advantages of both MRT and ZF precoders.

As explained in [9] and [42], the TPE precoding matrix is

$$W_{TPE} = \sum_{j=0}^{J-1} w_j (H^T H^*)^j H^T \tag{15}$$

where w_j is the coefficient of the precoder polynomial of order J. An appropriate selection of J ensures a smooth performance transition between the conventional low-complexity MRT when $J = 1$ and the rising complexity RZF when $J = min(M; K)$.

In terms of system efficiency and implementation complexity, TPE has the following advantages: [9], [42], [49]: i) The computational delay is reduced because computations are parallelized across multiple processing cores and operations are distributed uniformly over time, ii) the parameter J can be instantly customized to the available hardware, and iii) there is no need to compute the precoding matrix in advance (which leaves more channel uses for data transmission).

5) PHASED ZF PRECODING TECHNIQUE

When a large number of antennas are installed at the basestation to enable high-speed communications, hardware complexity becomes more of an issue. In [11], phased ZF (PZF) is envisioned as a low-complexity ZF-based precoding technique that takes full advantage of the low-dimensional RF chain to confront this fundamental challenge. It only regulates the phase in the RF domain, allowing for low-dimensional baseband ZF precoding. The PZF downlink transmission method consists of two stages: baseband (W) and RF joint processing (F). The baseband precoder W (dimensions K x K) modifies the amplitudes and phases of incoming complex symbols, whereas the RF precoder F (dimensions M X K) controls the phases of the converted RF signal.Prominently, amplitude and phase adjustments are possible for the baseband precoder "W," and only phase adjustments with variable phase shifters and combiners are possible for the RF precoder "F". As a result, each F entry is standardized to achieve $|F_{i,j}| = \frac{1}{M}$, where $|F_{i,j}|$ represents the magnitude of the $(i, j)^{th}$ element of F. If $F_{i,j}$ is the $(i, j)^{th}$ element of F, then, we perform RF precoding as described in [11].

$$F_{i,j} = \frac{1}{\sqrt{M}} e^{j\phi_{i,j}} \tag{16}$$

where, $\phi_{i,j}$, is the phase of the $(i, j)^{th}$ conjugate transpose of the composite downlink channel, i.e. $[h_1, \ldots, h_K]$. At the baseband precoder, we notice a comparable channel $H_{eq} = HF$ of a low dimension K $\times$ K where $H = [h_1, \ldots, h_K]^H$ is the composite downlink channel. Thus, the multi-stream baseband channel is applied to H_{eq}, where PZF precoding is performed as [9], [11]

$$W_{PZF} = H_{eq}^H (H_{eq} H_{eq}^H)^{-1} \lambda \tag{17}$$

where $\lambda \in \mathbb{R}^{K \times K}$ is a positive diagonal matrix for column power normalization. It could be seen that $W_{PZF} \in \mathbb{C}^{K \times K}$, whose number of rows is intensely reduced from M to K. Besides, in [9] it is also proofed that the spectral efficiency of a PZF precoder is closely upper bounded by that of a ZF precoder.

By extracting phases of the conjugate transpose of the cumulative downlink channel from the basestation to multiple users, the PZF precoding deploys phase-only control in the RF domain. It intends to coordinate the phases of channel elements to obtain the comprehensive array gain provided by excessive antennas in massive MIMO systems. PZF approaches ZF precoding performance, although it is practically impossible due to the requirement of supporting each antenna with its own RF chain [11].

6) ADVANTAGES AND DISADVANTAGES OF LINEAR PRECODING TECHNIQUES

Table 1 summarizes the benefits and drawbacks of linear precoding methods [9], [11], [38], [43]–[45], [48], [49], [51].

7) MASSIVE MIMO LINEAR PRECODING PERFORMANCE EVALUATION

From a practical point of view, the criteria for analyzing the performance of massive MIMO linear precoding schemes are: bit error rate, signal to noise ratio, sum rate, and energy efficiency [52].

The result in Fig.5 describes the sum rate and signal to noise ratio for zero-forcing, regularized zero-forcing, and maximum ratio transmission with $M = 64$ and $K = 8$. The bandwidth efficiency of MRT is better than ZF in the region of lower values and vice versa in the higher SNR values. As the signal-to-noise ratio is increased, a higher impact of interference on performance is observed. This suggests that the MRT precoding method is a viable solution when transmit

TABLE 1. Advantages and disadvantages of linear precoding techniques.

Linear Precoding	Advantages	Disadvantages
ZF	•Compared with nonlinear precoding, the signal processing is simple •Better performance at high SNR's •The SINR can be made according to preference by increasing the transmit power •Takes into account inter-user-interference •More power efficient •It functions better under noise free channel	•Under noise limited conditions, it works badly •More power consumption •Noise amplification and power penality for highly correlated channel •Higher implementation complexity •Unable to support too many users
MRT	•Near optimal performance if there are more BS antennas than users •Computational complexity is lower because of the absence of channel matrix inversion •Is preferred for high mobility scenario •Better performance at lower SNRs	•Unable to achieve full diversity at high SE •Having lower achievable rate in the case of less BS antennas •Suffering from error floors for positive multiplexing gains
RZF	•It is optimal if SINR and average channel attenuation ratio is the same for all users •The sum rate performance of RZF is better than ZF and MRT	•Complexity for large matrix inversion •Suffering from error floors for all multiplexing gains
TPE	•Balances complexity and throughput •Replaces matrices inversion by polynomial expansion •More suitable for real-time hardware implementation	•Lower order polynomials have limited performance •Higher performance requires considerably more hardware
PZF	•Low-complexity facilitating multi-stream processing •Its performanc is alsmost the same as ZF •Large power gain	•SE being tightly upper bounded by ZF precoder

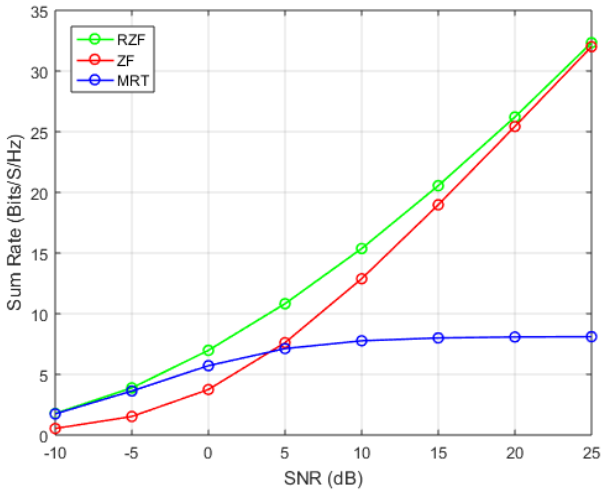

FIGURE 5. Sum rate versus signal to noise ratio.

power is minimal. At high SNR levels, the sum-rate achieved by ZF and RZF precoding schemes is significantly greater than that of MRT. The total rate with ZF and RZF precoding schemes increases exponentially as the SNR increases because these precoding strategies promote IUI and noise cancellation. RZF's achievable rate outperforms ZF and MRT across the entire spectrum of SNRs.

Figure 6 depicts the spectral efficiency with the BS antennas at 10dB SNR. The study implies that increasing the number of basestation antennas significantly improves spectral efficiency, implying that the large-scale antenna significantly benefits the sum rate. Furthermore, the spectral efficiency with ZF and RZF precoding schemes is more than double that of MRT. The result also shows that the overall achievable rate with RZF precoding is higher than with ZF precoding, indicating that the RZF precoding system is the preferred solution for the less computationally complex massive MIMO approach.

The simulation result in Fig. 7 shows when BER decreases, SNR increases which indicate the signal is stronger than noise. Because the ZF precoder removes interference, it outperforms the RZF precoding technique; however, ZF implementation is extremely complex when compared to other

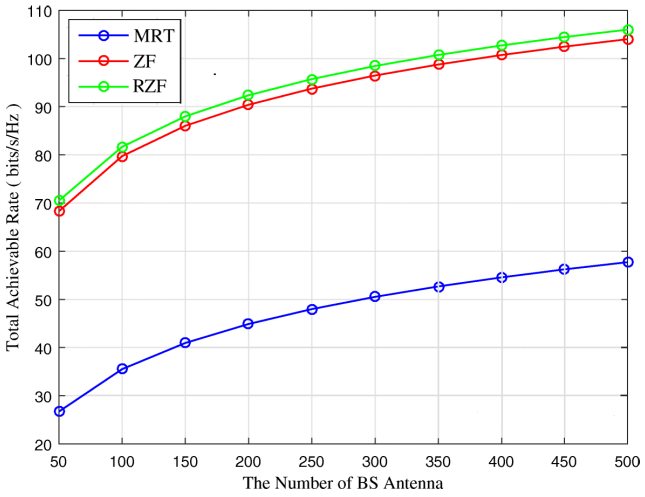

FIGURE 6. Achievable rate versus basestation antennas.

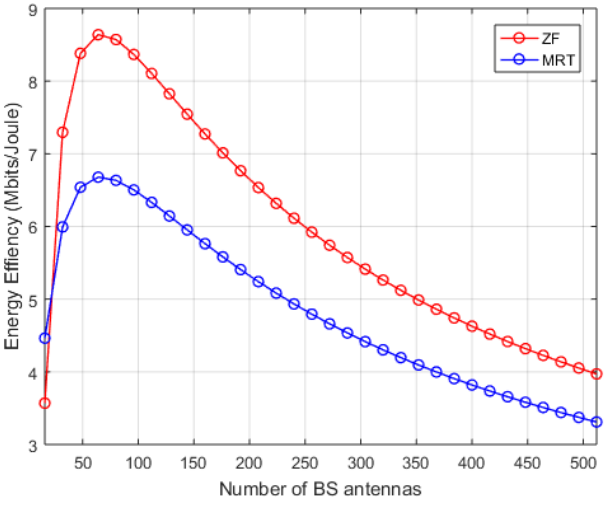

FIGURE 8. Energy efficiency versus number of base station antennas for ZF and MRT.

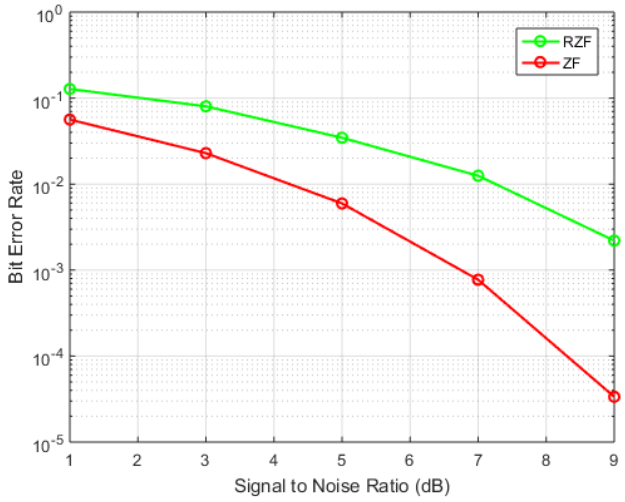

FIGURE 7. Bit error rate versus signal to noise ratio.

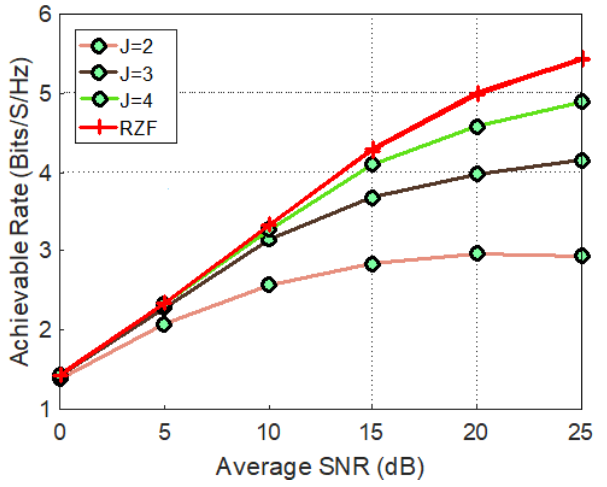

FIGURE 9. Achievable sum rates of TPE and RZF precoding techniques.

linear precoding techniques. On the other hand, due to the consideration of noise terms, RZF's performance is much less than ZF's, especially at higher values of SNR.

The result in Figure 8 describes the increment of energy efficiency until some number of basestation antennas and then decreases with the number of antennas. It is because of the increment in the overall circuit power consumption in the BS, as shown in equation (10). This depicts that as BS antennas grow high, it results in lower energy efficiency. Thus, while increasing the number of basestation antennas may help to reduce network transmitting power, it decreases the energy efficiency due to an increase in internal power consumption. Therefore, to obtain the optimal number of basestation antennas while also providing maximum spectral and energy efficiency, a design trade-off is required.

The achievable sum rate of TPE and RZF precoding techniques against average received SNR is depicted in Fig. 9.The simulation employs M = 512 antennas and K = 128 users because this is the regime in which TPE performs poorly and precoding complexity becomes a problem [50]. With increasing J value, the achievable rate of TPE approaches that of RZF. As the value of J increases, TPE efficiency approaches RZF even though it enhances hardware complexity. TPE-based schemes approach the sum rate and minimum user rate of the optimal RZF precoder even for a small number of matrix polynomial terms. TPE precoding, of course, never outperforms RZF performance, which is natural since TPE precoding is an approximation of RZF. Even for a small number of matrix polynomial terms, TPE-based schemes approach the sum rate and minimum user rate of the optimal RZF precoder.

8) COMPARISON OF MASSIVE MIMO LINEAR PRECODING TECHNIQUES

The comparison in Table 2 considers the parameters bit error rate, signal-to-noise ratio, spectral efficiency, and energy

TABLE 2. Summery of linear precoders performance comparison.

Performance Parameters	Linear Precoders				
	ZF	MRT	RZF	TPE	PZF
Bit Error Rate	Lowest	Lowest	High	Medium	Low
Signal to Noise Ratio	Higher	Highest	High	Low	Medium
Achievable Sum Rate	Medium	Low	Higher	High	Highest
Energy Efficiency	Medium	Low	High	Higher	Highest
Complexity	Full Complex	Lowest	Complex	Low Complex	Lower
Performance	High	Low	Optimal	Tends to RZF	Approaches ZF

efficiency of the linear precoder approaches [9], [11], [38], [43]–[45], [48], [49], [51].

9) NON-LINEAR PRECODING TECHNIQUES

Extensive research has been conducted to identify the best precoding method for massive MIMO systems, which accomplishes a desirable throughput performance while reducing complexity. However, even though the linear precoders have the advantage of low complexity, their insufficiency in precoding accuracy cannot be neglected [61]. The optimum linear signal precoder is zero-forcing that attains optimal performance and has a robust potential to decrease the error probability. It has an illegitimate complexity due to a large number of antennas in massive MIMO systems. Unfortunately, ZF comprises a matrix inversion during processing. This makes ZF precoding methods computationally inefficient for a large number of antennas [24]. In general, linear precoding has good performance when the channel correlation between the users is low, but this performance degrades severely when the channel correlation between the users is high. On the other hand, non-linear precoding schemes are usually more robust to channel correlation between UEs and have the potential to significantly enhance the massive MIMO performance in 5G. This makes non-linear precoding be identified as a candidate massive MIMO precoding scheme for 5G new radio (NR) [62].

This section, therefore, introduces non-linear massive MIMO signal precoding schemes with their basic facts, comparisons, and performance analysis.

10) DIRTY PAPER CODING

The DPC algorithm states that if the interference is known at the transmitter side, the capacity of the theoretical channel is offered, and the interference can be canceled [24]. It is a technique of canceling known interference at the transmitter. For a channel

$$Y = X + S + W \tag{18}$$

where $X = [X1, X2 \ldots XL]$ is the sequence of power limited transmitted symbols with power $P_x = \frac{E[||X||^2]}{L}$, $S = [S_1, S_2 \ldots S_L]$ is a known interference sequence at the transmitter and W is the white Gaussian noise vector with $P_w = \frac{E[||W||^2]}{L}$, the capacity is the same as if we have no interference S, i.e, $Y = X + W$ [63]. When the precoding matrix in multi-user MIMO systems is designed for the n^{th} received terminal, interference from the first up to the $(n-1)^{th}$ received terminals are considered to be nullified. Furthermore, the DPC provides exceptional performance without requiring additional power on the transmission side or sharing CSI with the receiver side. In the other hand, it is unrealistic because DPC requires an infinite number of codewords and complex signal processing [64]. The DPC has been proposed to provide the ideal downlink sum rate for massive MIMO systems, which is based on the idea that the sum rate of a system is equal to the sum rate of a free-interference system when the interference is known at the transmitter side. The DPC sum-rate capacity is given as [24]

$$C = max_w \log_2 det(I_N + H^* W H^T) \tag{19}$$

where W is a N X N diagonal power allocation matrix, and $\sum diag(W) = 1$.

11) TOMLINSON HARASHIMA

Tomlinson and Harashima independently introduced a new precoding technique to mitigate ISI. The structure they presented is referred to as Tomlinson-Harashima Precoder (THP). There are other precoding structures that are combined for better performance [65]. To achieve a lower complexity implementation and provide more flexibility for various types of CSI feedback while offering substantial performance gain, the conventional linear precoders are combined with a Tomlinson-Harashima precoder (THP) [62]. For example, the research work mentioned in [65] showed that zero-forcing THP (ZF-THP) achieves the maximum possible mutual information at high SNR for pulse amplitude modulation (PAM) inputs. The authors in [9] introduced the MMSE-THP and compared its output SNR with that of the ZF-THP. This correspondence quantifies the loss capacity of ZF-THP and MMSE-THP for any given ISI channel and an AWGN variance. For channels with severe ISI, the MMSE-THP can provide a significant performance improvement over the ZF-THP from low to mid-range SNR. However, when the SNR becomes sufficiently large, the two techniques become identical.

TH precoding is a combination of the DPC, and modulo arithmetic [66]. It is a suboptimal implementation of the DPC, which has an equalization process proposed to repeal the ISI [24]. TH can also be used to clear sub-channel interference in MIMO systems. The TH precoding has more complexity when compared to linear precoding algorithms but effectively avoids noise amplification.

A standard THP algorithm implements three filters, the feedback filter $B \in \mathbb{C}^{N_t \times K}$, the feedforward filter $F \in \mathbb{C}^{N_t \times N_t}$ and the scaling matrix $G \in \mathbb{C}^{N_t \times N_t}$ [67], [68]. The feedback filter deals with the multiuser interference by successively subtracting the interference from the current symbol. The matrix B has a lower triangular structure, where as the feed-forward filter enforces the spatial causality. The scaling

filter assigns a coefficient or weight to each stream of data, which means G is a diagonal matrix.

The two general THP structures in the literature are the centralized THP (cTHP) and the decentralized THP (dTHP) [69]. Their difference is the scaling matrix G is placed at the transmitter for the cTHP, whereas for dTHP the same matrix is located at the receiver. The TH algorithm is carried done by decomposing H^T via LQ decomposition to the multiplying of a lower triangular matrix $\mathbf{L}$ and a unitary matrix $\mathbf{Q}$ as $H^T = LQ$ [24], [69]. Thus, the THP filters are expressed as:

$$\begin{aligned} F &= Q^H, \\ G &= diag(l_{11}, l_{22}, \ldots, l_{KK})^{-1} \\ B^d &= GL \\ B^c &= LG \end{aligned} \tag{20}$$

where B_d and B_c correspond to the feedback filter for dTHP and cTHP, respectively.

12) VECTOR PERTURBATION

Vector perturbation (VP) is designed to present an easy encoding technique without using DPC. It is also viewed as a generalized TH scheme [24]. When compared to the DPC, TH provides a full diversity order at an even lower computational complexity. The vector of perturbed data is precoded by a linear front-end precoder to minimize the unscaled transmitted power [70]. The VP algorithm's perturbation operation requires the linear front-end precoding system to obtain the perturbing vector of the signal to be sent to all users. This indicates that these two processes must be performed concurrently. The TH, on the other hand, selects the scalar integer offset to be used in the transmitter sequentially and does not perform nearly as well as the VP algorithm choice. Instead of modifying the inverse process, the VP technique modifies the transmitted data by aligning the data symbols at the transmitter to the eigenvalues of the inverse channel matrix on an instantaneous basis [24]. The generator matrix of the lattice is the pseudo-inverse of the channel matrix and its regularized version. This is accomplished by discretely inserting a scalar integer vector offset at the transmitter, which results in interference cancellation at the receiver via a modulo arithmetic operation [71]. This confirms that the VP provides an alternative non-linear coding method capable of achieving near DPC performance with less complexity [72].

Since VP precoding employs a channel inversion matrix and applies a perturbation on transmitted symbols, the transmitted signal is [73].

$$x = \sqrt{\frac{P}{\beta}} \hat{H}^{\dagger}(u + \tau l^*) \tag{21}$$

where $\hat{H}^{\dagger}$ is the Moore-Penrose generalized inverse of matrix $\hat{H}$ which is the estimate of the channel matrix at the transmitter, $u_k \in \mathbb{C}^{M\times 1}$ is data symbol vector

$$\beta = \| \hat{H}^{\dagger}(u + \tau l^*) \|^2 \tag{22}$$

is the transmit power scaling factor so that $\mathbb{E} \| x \|^2 = P$ and $l^* \in \mathbb{C}^{M\times 1}$ is the selected perturbation vector with integer entries. $\tau = 2|c|_{max} + \Delta$ is the absolute value of the constellation symbol with the maximum magnitude and Δ denotes the minimum Euclidean distance between constellation symbols. Thus, the received symbol is [24], [73]

$$x = \sqrt{\frac{P}{\beta}} \hat{H}\hat{H}^{\dagger}(u + \tau v^{\star} + w) \tag{23}$$

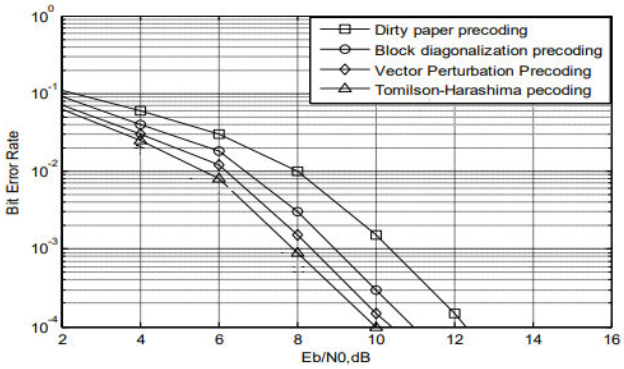

FIGURE 10. Non-linear precoding techniques BER performance [74].

Block diagonalization is one of the methods for linear precoding in which power optimization is done for a number of antennas instead of a single antenna element. Thus, Fig. in 10 compares block diagonalization with non-linear precoding techniques. In particular, Fig. 10 indicates that the BER plot is a monotonically decreasing function of SNR. It almost declines to lO^{-3} at SNR of 11dB, 9dB, 8.5dB and 8dB with DPC, BD, VP and TH precoding techniques respectively. Except for its complexity, DPC has a much lower bit error rate than its counterparts.

The summary of non-linear precoding techniques with their respective advantages and disadvantages is shown in Table 3 [24], [62], [64]–[66], [69]–[71].

III. MILLIMETER WAVE COMMUNICATIONS

The ultimate focus of the 5G system is to enhance data rates by a factor of 1,000 over the preceding 4G cellular systems [75]. Analogously, it is plausible to shift to frequency bands that have not been used in the previous wireless communication era. In this reference, the mmWave is the insanely high-frequency range from 30-300 GHz in the electromagnetic spectrum, with a corresponding wavelength of 1-10 mm [76]. When compared to today's wireless communications, mmWave systems get the following advantages: extensive bandwidth and capacity, narrow beams, small element size, low probability-of-interception, and directivity [77]. Even though the mmWave system has these advantages, its signal transmissions face profound challenges such as aggressive path loss, penetration loss, power consumption, shadowing, blockage, hardware impairment, and so on [78]. In this regard, the free-space reference-distance path-loss (P_L) with

TABLE 3. Comparison of the advantages and disadvantages of the non-linear precoding techniques.

Non-Linear Precoding	Advantages	Disadvantages
DPC	•Provides high sum rate capacity •Offers the best performance than other precodings •Cancels the known interference	•High computational complexity which leads to unfavourable complexity specially in massive MIMO •Costly due to its high complexity and power consumption
TH	•Close capacity performance •Effectively reduces noise amplification •Has modulo operation to cancel ISI, limit transmit power and avoid non linear distortion •Acts as spatial equalization	•Computational complexity is high for large ratio of M/N •Has sensitivity to CSI inaccuracies •Transmit power may increase due to having SIC process •The use of modulo operation leads to loss of performance as shaping and power losses •Suffers diversity penality
VP	•Improves channel inversion performance and offers near capacity performance •Less complex precoding than DPC •Reduces transmit power by purturbing the signal vector with additional vector •The variables of power-scaled transmitters could be data dependent •Obtains the optimal purturbation vector by solving a minimum distance type problem	•Power-scaled factors are data dependent that has insignificant overhead during transmission •Dynamic range comparable to the power-scaled factor •Does not employ adaptive modulation •Performance degrades in the limited feedback scenarios •It suffers from the CSI imperfections and inaccurate power-scaled factors

respect to [79] is:

$$P_L[dB](d) = 20\log(\frac{4\pi d_o}{\lambda}) + 10n\log(\frac{d}{d_o}) + X \quad (24)$$

where, d_o, d and n are the reference distance, transmitter-receiver separation distance and path-loss exponent respectively. X is a shadow fading term. The distinction among today's wireless network and mmWave systems is with respect to the free-space path-loss model in equation (24) that results in the variation with wavelength, as seen in the denominator of the first term of the path-loss expression.

The use of many antenna arrays to resolve the poor path-loss scenario in any of those bands is crucial to function in the mmWave bands. That is, the size of the antenna array is affected by the number of individual antenna elements, which is impacted by the device's center frequency. As the carrier frequency increases, the size of the antenna array decreases, making it possible from a form-factor standpoint to have mmWave arrays with hundreds of elements [78].

A. MILLIMETER WAVE SYSTEM MODEL

Taking into account the downlink multi-user millimeter wave MIMO communication scheme with the basestation serving K users concurrently (Fig. 11), the basestation is equipped with N_{RF} RF chains and N_t transmit antennas. Every mobile station is equipped with N_r receiving antennas. The transmitter communicates with each mobile station via a single stream that makes the total number of transmitted streams to be $N_S = U$. The amount of mobile devices does not exceed that of BS RF chains, i.e., $N_S \leq N_{RF}$. The CSI is assumed

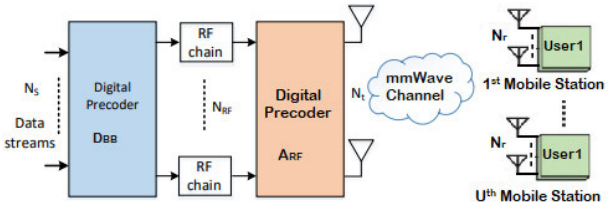

FIGURE 11. Hybrid precoding model in mmWave communications [80].

to be perfect at the BS. The precoder is the hybrid with the analog and baseband precoders.

The spatial multiplexing gain of hybrid precoding is limited by $N_{RF} \leq N_t$. The BS applies an $N_{RF} \times N_S$ digital precoding matrix $D_{BB} = [d_{BB}^1, d_{BB}^2, \ldots, f_{BB}^k]$ where $f_{BB}^k (k \in [1, 2, \ldots, K])$ is an $N_{RF} \times 1$ vector, with an $N_t \times N_{RF}$ RF precoding matrix. The analog precoding matrix (A_{RF}) is used as analog phase shifter and its parameters are constant modules which is required to satisfy $|A_{RF}[i, j]| = \frac{1}{\sqrt{N_t}} \forall_i = 1, 2, \ldots, N_t, j = 1, 2, \ldots, N_{RF}$ which is the $(i, j)^{th}$ element of analog precoding matrix. Thus, the transmitted part of the signal is given by the expression

$$x = A_{RF} D_{BB} s \quad (25)$$

where $s = [s_1, s_2, \ldots, s_k]^T$, transmitted message $(N_s \times 1 (k \times 1))$ vector and normalized as $\mathbb{E}[ss^H] = \frac{1}{(N_s)} I_k$. The hybrid beamforming matrix total power constraint of is normalized as D_{BB} as $\|D_{BB} A_{RF}\|^2 = N_S$. Therefore, the received

signal y_k detected by the k^{th} mobile station (MS) is [81]

$$y_k = \sqrt{\rho} H_k \sum_{n=1}^{K} A_{RF} D_{BB} s_n + n_k \tag{26}$$

where $H_k \in \mathbb{C}^{N_r \times N_t}$ is the channel between the *BS* and the k^{th} user. $n_k \in \mathcal{CN}(0, \sigma^2)$ and ρ are the Gaussian noise vector and average received power respectively.

B. MILLIMETER WAVE CHANNEL MODEL

Millimeter-wave channels are predicted to have limited scattering. In mmWave, every scatterer is expected to add a single path of propagation between the transmission and reception [82]. As far as mmWave channel modeling is concerned, an extended Saleh-Valenzuela model is commonly used. This model is used in our research because of the high antenna correlation, and low spatial selectivity [83]. The use of a geometric channel model with L_k scatters is applied for the channel of the k^{th} user. In this model, the k^{th} user channel matrix $H_k \in \mathbb{C}^{N_r \times N_t}$ can be expressed as

$$H_k = \sqrt{\frac{N_t N_r}{L_k}} \sum_{l=1}^{L_k} \alpha_{k,l} a_{k,r}(\theta_{k,l}) a_{k,t}^H(\phi_{k,l}) \tag{27}$$

where $\alpha_{k,l}$ represents complex gain of l^{th} path. It contains path-loss with $\mathbb{E}|\alpha_{k,l}|^2 = \beta$ (normalization constant). $\phi_l \in [0, 2\pi]$ and $\theta_l \in [0, 2\pi]$ are l^{th} azimuth angles of *AoDs* and *AoAs*) of transmitter and receiver respectively. $a_{k,t}(\phi_{k,l})$ and $a_{k,r}(\theta_{k,l})$ denotes antenna array response vector of the transmitter and receiver respectively with their respective antenna array response vector of k^{th} *MS*. L_k is the amount of propagation paths for the channel of k^{th} *MS* with $l \in [1, 2, \ldots, L_k]$. The BS and MS get knowledge about the structure of their antenna arrays. If a ULA is considered, the array response vectors at the transmitter, $a_t(\phi)_l$ is defined as

$$a_t(\phi)_l = \frac{1}{\sqrt{N}} [1, e^{j(\frac{2\pi}{\lambda})d \sin(\phi_l)}, \ldots, e^{j(N_t-1)(\frac{2\pi}{\lambda})d \sin(\phi_l)}]^T \tag{28}$$

On the other hand, the receiver's array response vectors is

$$a_r(\theta_l) = \frac{1}{\sqrt{N}} [1, e^{j\frac{2\pi}{\lambda} d(xsin(\phi)sin(\theta)+ycos(\theta))}, \ldots \times e^{j\frac{\pi}{\lambda} d((W_1-1)sin(\phi)sin(\theta)+(W_2-1)cos(\theta))}]^T \tag{29}$$

where λ denotes the wavelength, d is the spacing between two antennas (commonly, $d = \frac{\lambda}{2}$), ϕ, = azimuth angle, θ = elevation angle and N = antenna elements number.

IV. MILLIMETERWAVE-MASSIVE MIMO COMMUNICATIONS

Massive MIMO is one of the technologies to achieve 5G requirements. It has plenty of antennas at the BS which serve multiple users simultaneously. In addition, the extra antennas help to focus energy into a smaller region of space, which makes the network more efficient in throughput, spectrum utilization, and energy consumption. MmWave communication is another technology with a large amount of unused bandwidth to support millions of devices at once [3]. Since the mmWave frequencies are highly directional as compared to lower frequencies, they can precisely handle large antenna arrays during the transmission and reception process with a beamforming strategy [3]. On the contrary, it is challenging for long-range wireless communication because of the huge path loss in the mmWave band. To compensate for the path loss and extend the coverage, massive MIMO antennas are supposed to be deployed in mmWave communication systems. Thus, it is natural to merge massive MIMO and mmWave, which dramatically improves wireless access, throughput or the quality of service in 5G cellular system [4]. This technological combination of massive MIMO and mmWave systems has given birth to mmWave-massive MIMO [84] which brings an opportunity to support a plethora of high-speed services for bandwidth-hungry applications. That is, mmWave massive MIMO has the capacity to boost user throughput, mobile network capacity, spectral and energy efficiency [85].

Figure 12 depicts the setup of three technologies: ultra-dense networks (UDNs), mmWave, and massive MIMO, which have a mutually beneficial relationship. The system is a HetNet made up of macro-cells and small cells *BSs*, all with massive MIMO and mmWave communication capacities.

The barriers that must be overcome so as to reap the benefits of mmWave-massive MIMO technology are [19]: beamforming, waveforms, channel modeling, multiple access schemes, antenna and RF transceiver architecture design, etc.

A. MILLIMETERWAVE-MASSIVE MIMO BEAMFORMING

Beamforming is employed to further enhance the spectral efficiency of massive MIMO communication systems [21]. It improves the performance of wireless systems by employing the interference cancellation concept. Beamforming is a method of preprocessing signals prior to transmission in order to optimize network throughput and reliability in a MIMO system by taking advantage of the available spatial degrees of freedom [86]. It also offers a high beam gain for wireless systems, allowing them to overcome the severe path loss challenges posed by mmWave as well as hardware limitations in massive MIMO schemes [87]. Basically, the three categories of beamforming architectures are analog, digital, and hybrid beamforming architectures [39], [40].

1) ANALOG BEAMFORMING

In analog systems, an RF chain with several analog phase shifters is used to transmit a single data stream (Fig. 13). Spatial processing is performed within the analog portion of the transceiver [21]. The transmit and receive array processing is carried out using RF components capable of phase shifting and possibly gain adjustment [88]. It is used to control the phases of original signals to achieve the maximal antenna array gain and effective SNR. Analog beamforming has a

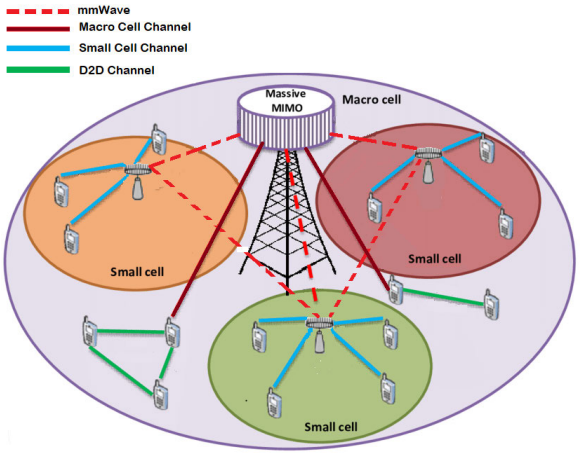

FIGURE 12. Network architecture for mmWave-massive MIMO for 5G [41].

simpler hardware structure that makes it easier to implement. On the other hand, it has poor antenna gain and suffers from major performance loss because it can only regulate the phases of the transmit signals but not their amplitudes. As a result, it is not practically implemented in massive MIMO and mmWave cellular technologies [41].

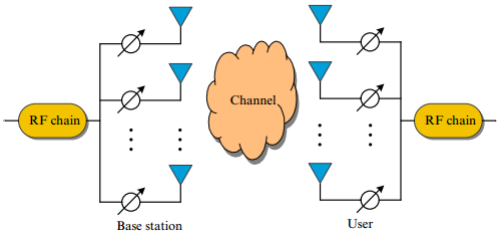

FIGURE 13. Analog beamforming architecture for mmWave massive MIMO system [41].

2) DIGITAL BEAMFORMING

In conventional MIMO systems, beamforming architectures are realized typically in the digital domain, where all the signal processing techniques are performed at baseband, as illustrated in Fig. 14. The digital baseband precoders allow for independent tuning of the magnitude and phase values of the transmit signals of each antenna element [89]. Its key concept is to monitor both the phases and amplitudes of the original signals to detect and eliminate interferences beforehand. In the case of digital beamforming, each antenna element requires a dedicated RF transceiver, which is prohibitively expensive. It is possible, for example, to use different precoding weights for each sub-carrier of mmWave massive MIMO systems, allowing one to tackle more difficult non-line-of-sight (NLOS) scenarios.

Fully digital beamforming necessitates a full RF chain for each antenna. This results in unviable complexity, power consumption and cost [22], [90] specially at high frequencies (mmWave range) that employs large number of antennas [40].

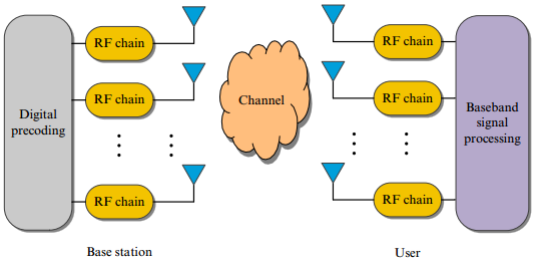

FIGURE 14. Digital beamforming architecture for mmWave massive MIMO system [41].

3) HYBRID ANALOG-DIGITAL BEAMFORMING

It is a type of architecture that combines the advantages of both analog and digital beamforming schemes [88]. That is, a portion of the processing is with a reduced number of RF chains in the baseband and the remaining part in the RF band. Such topologies (Fig. 15) offer an optimal balance between analog and digital processing [21]. This makes a hybrid beamforming architecture an appealing option for mmWave-massive MIMO systems, particularly as the number of antennas grows large [91]. Since the general design principle of hybrid analog-digital precoding is to approximate the full-digital precoder, many works in literatures [92]–[94] have proposed hybrid precoding designs that achieves the optimal full-digital precoding performance. Besides, the hybrid precoding design is known to improve the spectral and energy efficiency of the system significantly [89].

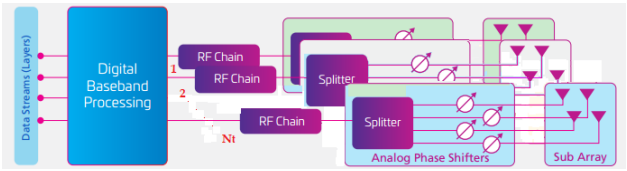

FIGURE 15. Hybrid analog-digital beamforming.

B. SYSTEM MODEL

1) MILLIMETER WAVE-MASSIVE MIMO BEAMFORMING MODEL

We consider a single-user mmWave-massive MIMO system with hybrid analog/digtal beamforming, where the BS utilizes N antennas to concurrently transmit N_s data streams to a user with M antennas (Fig. 11). Thus, to enable multi stream transmission, the BS has N_{RF} RF chains with $N_s \leq N_{RF} \leq N$. The BS uses $N_{RF} \times N_s$ digital precoder D_{BB} using its N_{RF} RF chains, followed by an $N \times N_{RF}$ RF analog beamformer A_{RF} using analog system. Then, the transmitted signal is [19], [95]

$$x = A_{RF} D_{BB} s \tag{30}$$

where s is $N_s \times 1$ original signal vector before precoding with normalized power as $\mathbb{E}(ss^H) = (\frac{1}{N_s})IN_s$. Therefore, for

a narrow band system the received signal vector y of size $M \times 1$ is

$$y = \sqrt{\rho} H A_{RF} D_{BB} s + n \tag{31}$$

To enable precoding, we assume the channel matrix H is be perfectly known at both the transmitter and receiver.

For a multi-user hybrid analog-digital beamforming system, a mmWave-massive MIMO with BS antennas N and N_{RF} RF chains is considered such that $N_{RF} \leq N$ and K terminals each with M antennas and one RF chain. For a fully-connected hybrid architecture, the BS uses $K \times K$ digital precoder in the baseband $D_k = [D_1, D_2, \ldots D_k]$ followed by$N \times K$ analog precoder $A_k = [A_1, A_2, \ldots, A_K]$. The received signal vector r_k seen by k^{th} terminal after precoding is described as [81], [95]

$$r_k = H_k \sum_{n=1}^{K} A_n D_n s_n + n_k \tag{32}$$

At the receiver, it is combined with analog combiner w_k, where w_k is assumed to have the same constraints as the analog precoder A_k. In this scenario, the signal y_k is expressed as

$$y_k = w_K^H r_k = w_K^H H_k \sum_{n=1}^{K} A_n D_n s_n + w_K^H n_k \tag{33}$$

where W_K^H is the precoding matrix of the mmWave channel for user k. The goal of hybrid beamforming is to design the analog and digital precoders at the BS and the analog combiners at the MS to enhance the sum rate.

Accordingly, the received SINR for the k^{th} user is calculated from the expression

$$\gamma_k = \left(\frac{\frac{P}{K}|w_k^H H_k A D_u|^2}{\frac{P}{K}\sum_{n\neq k, n=1}^{K}|w_k^H H_k A D_u|^2 + \sigma_n^2}\right) \tag{34}$$

Finally, the system capacity R can be described as

$$R = B \sum_{k=1}^{K} \log_2(1 + \gamma_k) \tag{35}$$

where B denotes the bandwidth

2) MILLIMETER WAVE-MASSIVE MIMO POWER MODEL

A system's power consumption generally includes radiated power, circuit power, and power consumed for signal processing. In this scenario, the total power consumption is [96]

$$P_T = P_t + N_{RF} P_{RF} + N_S P_{PS} \tag{36}$$

where P_t denotes transmission power, P_{RF} describes the power consumed by RF chain and P_S is the power consumed by the power supply, N_{RF} and N_{PS} are the amount of RF chains and power supplies respectively.

3) ENERGY EFFICIENCY AND SPECTRAL EFFICIENCY MODEL

The energy efficiency metric is a "green communication" indicator. Essentially, paramount energy efficiency is acquired by exceptionally directing the radiated energy from the transmitter to focus into small regions where the active user equipment is found [97]. It is determined by the data rate over the mmWave links. It is calculated as the ratio of achievable rate to total power consumed by the system.

Spectral efficiency is proportional to transmit power. Increasing the signal power will improve the system's spectral efficiency. On the contrary, as signal power increases, so does system power consumption, which highly decreases the system's energy efficiency. That is, energy and spectral efficiencies are two trade-off parameters. Nevertheless, using large antenna arrays can improve spectral and energy efficiency compared to a single antenna system. The *SE* and *EE* of the downlink wireless system is [98]

$$\eta_{SE} = \frac{R}{B}, \quad \eta_{EE} = \frac{B\eta_{SE}}{P_T} \tag{37}$$

where, SE and EE are the spectral efficiency and energy efficiency respectively

C. PERFORMANCE OF MMWAVE-MASSIVE MIMO

The performance of mmWave-massive MIMO beamforming can be evaluated in terms of SNR, data rate, spectral efficiency, and energy efficiency, as it is also reported in our work [95].

Figure 16 shows that as the SNR increases, the spectral efficiency of the various beamforming schemes is enhanced to varying degrees. Different from other beamforming methods, the performance of fully digital beamforming is much higher than the others due to the dedicated RF chains for each antenna element. Because of the highest performance of digital beamforming, the remaining beamforming technologies are aimed at reaching it. Of course, the performance of hybrid beamforming is comparable to that of the full-digital system, though it uses fewer numbers of RF chains that result in lower power consumption and cost. The performance of analog precoding is the worst of all beamforming schemes since it only regulates the phase of the original signal in the analog domain.

A significant amount of gain is achieved by hybrid precoding over analog beamforming as the number of antennas increases, as shown in Fig.17). The difference in performance between hybrid and digital beamforming narrows as the number of BS antennas increments. For example, when the number of antennas is 75, the equivalent spectral efficiency of the hybrid and digital is 6.7 bps/Hz and 7 bps/Hz, respectively. When the number of antennas increases to 150, the equivalent spectral efficiency of hybrid and digital beamforming rises to 8 bps/Hz and 8.1 bps/Hz, respectively. That is, when the number of antennas increases from 75 to 150, the performance of the digital beamforming rises only 1.1 bps/Hz while that of hybrid beamforming increases by 1.3 bps/Hz. Since only the phases of the transmit signals are controlled but not

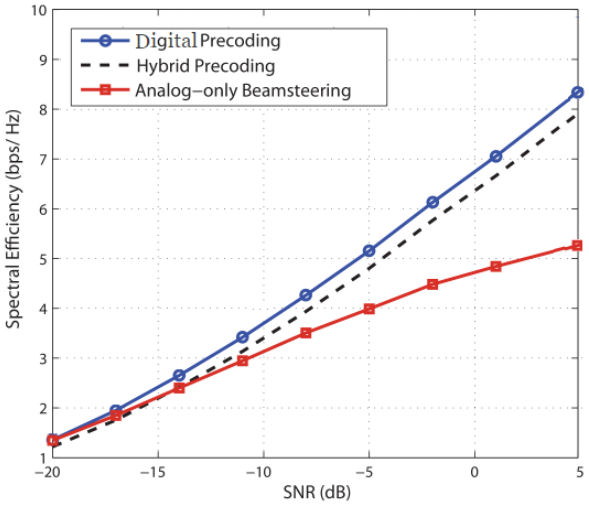

FIGURE 16. Spectral efficincy versus SNR for mmWave-massive MIMO system.

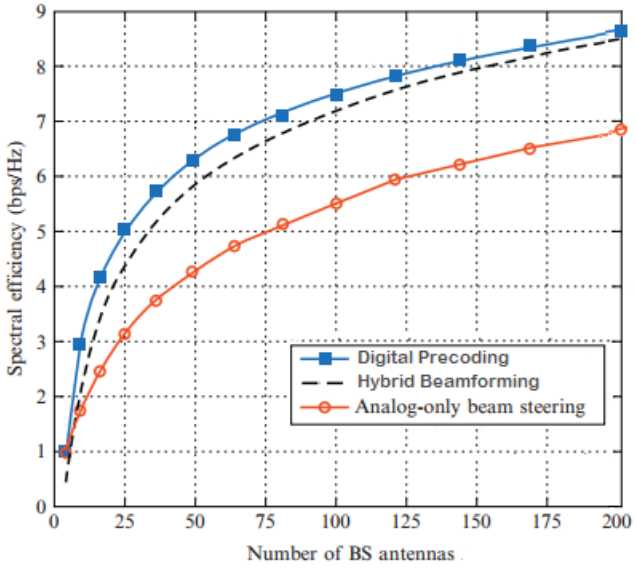

FIGURE 17. Spectral efficiency with *BS* antennas.

their amplitudes, the analog beamforming architecture has a non-negligible performance loss.

Figure 18 depicts the spectral efficiency with SNR for various linear precoding techniques. It is shown that hybrid linear precoding schemes outperform analog schemes over the entire SNR range. That is, the difference in performance widens as SNR rises. For instance, with SNR $= -15$ dB, the capacity gap between the linear precoders is less than 1 bps/Hz. But, when SNR $= 10$ dB, the capacity gap rises up to 5 bps/Hz. The performance of the analog precoding is the worst due to the analog constant modulus, which assures that only the phase characteristics are used, and the amplitude characteristics are not used at all.

As shown in Figure 19, as the number of antennas increases, so does the power consumption. This results in

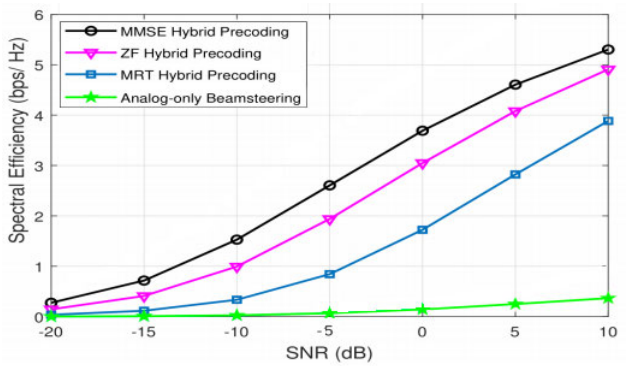

FIGURE 18. Spectral efficiency versus SNR with different precoding schemes.

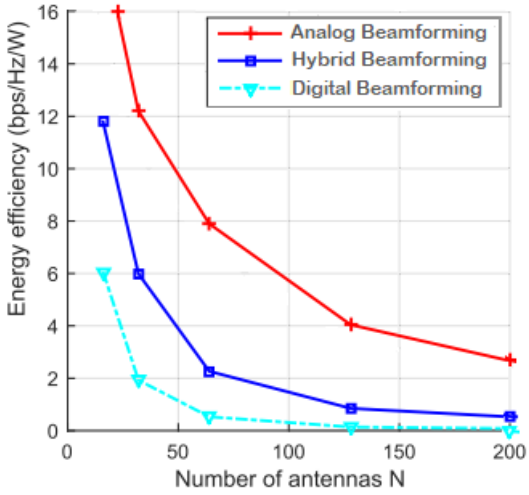

FIGURE 19. Energy efficiency versus number of base station antennas.

lower energy efficiency. This is due to the basestaion's overall circuit power consumption. That is, even though it helps to reduce the transmitting power for wireless communication by increasing the number of basestaion antennas, the energy efficiency declines because of the increase in internal power consumption. Consequently, the energy efficiency of digital precoding reduces as the number of antennas rises. On the contrary, hybrid beamforming has better energy efficiency than digital beamforming. The energy efficiency of analog beamforming is better than both digital and hybrid beamforming. However, due to the poor overall performance of analog beamforming, it is not used in real scenarios of mmWave massive MIMO communications.

D. HYBRID BEAMFORMING MODELING

Since the stochastic channel models require low computational complexity, they are a popular choice for mmWave system design [21], [27]. Thus, for analyzing the hybrid beamforming models, in this subsection, we use the "Saleh-Valenzuela" model [91]. In particular, we consider multiuser MIMO downlink communication systems, where precoding

is at the basestation side and combining is at the mobile station (Fig. 20).

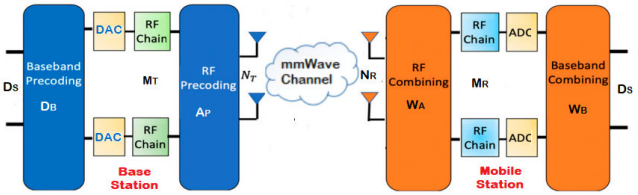

FIGURE 20. **Millimeter wave massive mimo model [3].**

The base station has M_T RF chains with transmit antennas N_T to support KD_S data streams, where K is the total amount of mobile stations. M_R is number of RF chains with receive antennas N_R. The transmitted data streams amount is limited by $KD_S \leq M_T \leq N_T$ and $D_S \leq M_R \leq N_R$ at BS and MS respectively. Transmitted symbols are processed by digital baseband precoder $B_B \in \mathbb{C}^{M_T \times KD_S}$ with analog RF precoder $A_P \in \mathbb{C}^{N_T \times M_T}$. Amplitude and phase changes could be done with D_B while A_P is used to modify phase only through analog phase shifters. Assuming that $|A_P(i,j)| = \frac{1}{\sqrt{N_T}}$, $|A_P(i,j)|$ is the amplitude of $(i,j)^{th}$ element of A_P and $||A_P D_B||_F^2 = KD_S$, where $||.||_F$ is the Frobenius norm that is used to adjust the total transmit power constraint. The received signal of multiuser MIMO to K^{th} MS user is

$$y_k = H_k A_P D_B s + n_k, \quad k = 1, 2, \ldots, K \tag{38}$$

$H_k \in \mathbb{C}^{N_R \times N_T}$ is the channel matrix of k^{th} MS, s is the total signal vectors of all MSs, with $s_k(D_S X 1)$ for each MS. Notably, $s = [s_1^T, s_2^T, \ldots, s_k^T]^T$, $[.]^T$ is transpose matrix and $\mathbb{E}[ss^H] = \frac{P}{KD_S} I_{KD_S} K$, where $\mathbb{E}[.]$ is the expectation, P is the average transmit power, I_{KD_s} is the identity matrix of size $KD_S X KD_S$ and $(.)^H$ is the conjugate transpose. n_k is additive Gaussian complex noise with zero mean and σ^2 variance. Then, the received signal after combining at k^{th} MS [99]

$$y_{\hat{k}} = W_{A(k)}^H W_{B(k)}^H H_k A_P D_B D_S + W_{A(k)}^H W_{B(k)}^H n_k, \quad k = 1, 2, \ldots, K \tag{39}$$

where $W_{A(k)}$ is analog RF combiner ($N_R X M_R$) and $W_{B(k)}$ is digital baseband combiner of K^{th} MS ($M_R X D_S$). Condition of $|W_{A(k)}| = \frac{1}{\sqrt{N_R}}$ can be applied for analog precoder for constant amplitude since $W_{A(K)}$ is implemented using phase shifters. Accordingly, its equivalent baseband channel model for each MS is

$$H_{\hat{k}} = W_{A(k)}^H H_k D_B, \quad k = 1, 2, \ldots, K \tag{40}$$

The processed received signal at the k^{th} MS (equation 39) could also be defined as

$$y_{\hat{k}} = W_{B(k)}^H H_{\hat{k}} D_{B(k)} s_k + \sum_{i=1, i \neq k}^{K} W_{B(k)}^H H_{\hat{k}} D_{B(i)} s_i + W_{A(k)}^H W_{B(k)}^H n_k, \quad k = 1, 2, \ldots, K \tag{41}$$

where the expression $\sum_{i=1, i \neq k}^{K} W_{B(k)}^H H_{\hat{k}} D_{B(i)} s_i$ denotes the interference and $W_{A(k)}^H W_{B(k)}^H n_k$ is noise signal. Finally, the total sum rate is [100]

$$R = \sum_{k=1}^{K} \log_2 \left| I_{D_s} + \frac{W_{B(k)}^H H_{\hat{k}} D_{B(k)} D_{B(k)}^H H_{\hat{k}}^H W_{B(k)}}{\sum_{i=1, i \neq}^{K} W_{B(k)}^H H_{\hat{k}} D_{B(i)} D_{B(i)}^H H_{\hat{k}}^H W_{B(k)}} \right| \tag{42}$$

The channel matrix for this work is set as $H = [H_1^T, H_2^T, \ldots, H_K^T]^T = [\sqrt{\beta_1}\dot{H}_1^T, \sqrt{\beta_2}\dot{H}_2^T, \ldots, \sqrt{\beta_K}\dot{H}_K^T]$ where $\sqrt{\beta_K}$ and $\dot{H}_K$ indicate the large scale path fading and normalized channel matrix respectively for the k^{th} MS, satisfying that $\mathbb{E}[\| H_k \|_F^2] = N_T N_R$.

To cancel inter-user-interference, hybrid precoders and combiners used for all K MSs are determined in hybrid beamforming block-diagonalization scheme that satisfies the optimal spectral efficiency of the MU-MIMO system for limited scattering mmWave channel environment [99]. In mmWave communications, large antenna arrays are applied to overcome the large free space path loss due to higher carrier frequency. A clustering channel model needs to be applied to the design to reduce the effect of scattering loss. That is, the mmWave downlink normalized channel for k^{th} MS is

$$\dot{H} = \sqrt{\frac{N_T N_R}{N_c N_L}} \sum_{i=1}^{N_c} \sum_{l=1}^{N_L} \alpha_{il}^k a_M^k(\theta_{il}^k) a_B^k(\phi_{il}^k)^H \tag{43}$$

where L is total scattered propagation paths in N_C clusters with N_L paths. To reflect the sparsity of the mmWave channel, both N_C and N_L should not be too large. α_{il}^k is channel complex gain of the l^{th} path in the i^{th} cluster with zero mean and σ^2 variance. θ_{il}^k is the k^{th} MS AoA and ϕ_{il}^k is the BS AoD. $a_M^k(\theta_{il}^k)$ and $a_B^k(\phi_{il}^k)$ are array response vectors of receive and transmit angles, respectively, assuming the angles are taken at azimuth direction only. Notably, standard deviations of AoD and AoA named as σ_ϕ^k and σ_θ^k respectively and it assumed to be constant for each i^{th} cluster. The stated channel parameters are used to obtain AoAs and AoDs for the channel matrix. The format of array response vectors changes when hiring different antenna structures, that is, either uniform linear array (ULA) or uniform planner array (UPA) [99], [101]. For ULA, the array response vector is

$$a_{ULA}(\theta) = \frac{1}{\sqrt{N}} \left[1, e^{j\frac{2\pi}{\lambda} d \sin(\theta)}, \ldots, e^{j(N-1)\frac{2\pi}{\lambda} d sin(\theta)}\right]^T \tag{44}$$

where λ is the wavelength of the carrier and d is the distance between neighboring antennas.

The UPA, which is employed by the BS and MSs, is given by [101].

$$a_{UPA}(\phi, \theta) = \frac{1}{\sqrt{N}} \Big[1, e^{j\frac{2\pi}{\lambda} d(n_z - 1)cos(\theta) + (n_y - 1)sin(\theta)sin(\phi)}, \ldots e^{j\frac{2\pi}{\lambda} d(n_z - 1)cos(\theta) + (n_y - 1)sin(\theta)sin(\phi)}\Big]^T \tag{45}$$

Fig. 21 shows the spectral efficiency for ULA and UPA. The channel that is considered is the mmWave channel with

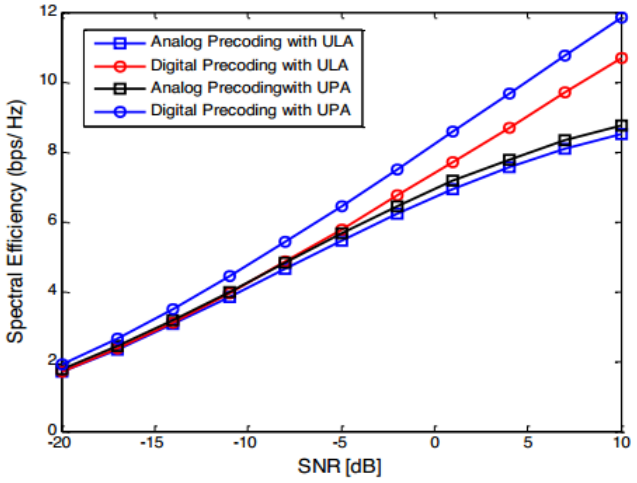

FIGURE 21. Spectral efficiency using ULA and UPA antenna arrays [101].

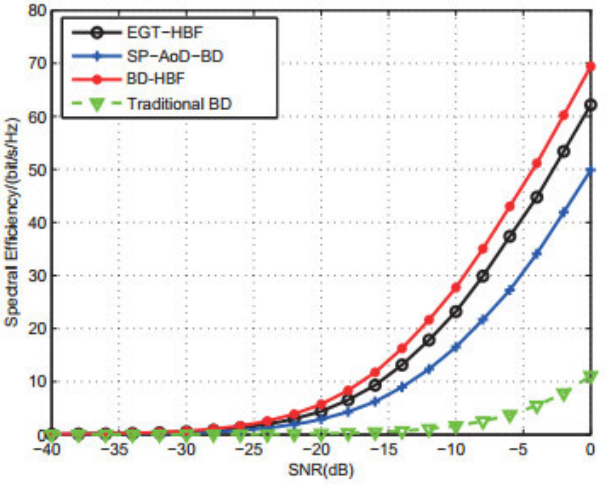

FIGURE 22. Spectral efficiency in sparse channel [99].

omnidirectional antennas at the users' terminals with regard to limited antenna elements. ULA takes into account only azimuth coverage, while UPA covers both azimuth and elevation coverage. As shown in Fig. 21, the simulation result shows that the performance of digital beamforming outshines its analog counterpart for higher SNRs.

E. HYBRID BEAMFORMING OPTIMIZATION

The constant-magnitude constraint makes optimizing hybrid beamforming difficult for the joint design of RF precoder/beamformer and baseband precoder/combiner. In [99], [100], iterative RF precoder/beamformer design algorithms have been proposed. These algorithms achieve a performance close to the digital beamforming. However, calculating the inverse matrix in each iteration makes the algorithm more complex. Thus, for the sake of simplicity, the optimization process is divided into two steps so as to separate the analog RF precoder/combiner design and the baseband digital precoder/combiner design. Designing the analog RF precoding/combining matrix is the first stage. The analog RF precoding matrix is designed by using equal gain transmission EGT). The aim here is to maximize the array gain while ignoring the interference. In the second stage, the extended BD algorithm and signal space can be used to eliminate inter-user-interference while enhancing signal power at the same time. Finally, precoding/combining will be performed with SVD to maximize the effective channel gain. Then, with the resulting effective channel, low dimensional baseband precoders could be effectively applied [108].

1) ANALOG BEAMFORMER DESIGN

The transmit and receive analog phase-only beamformers are $A_P \in \mathbb{C}^{N_R \times M_R}$ and $B_K \in \mathbb{C}^{M_R \times N_R}$ with the transmit and receive digital precoders $D_B \in \mathbb{C}^{M_R \times M_R}$ and $C_K \in \mathbb{C}^{M_R \times M_R}$ respectively. The analog beamformer F_r is designed to maximize the beamforming gain by employing equal gain combining

$$[F_r]_{i,j} = \frac{1}{\sqrt{N_R}} e^{j\phi_{i,j}} \tag{46}$$

where $\phi_{i,j}$ is the phase of $(i,j)^{th}$ element of H^H. The EGT precoding algorithm is only appropriate if $N_R = KN_U$ [99], [100].

2) DIGITAL BEAMFORMER DESIGN

We design C_k and D_B, using the BD approach [8], to cancel the inter-user-interference. The equivalent multiple access-phase channel with analog beamformer is $H_E = H_r[H_1, H_2, \ldots, H_K] = \hat{H}_r[\hat{H}_1, \hat{H}_2, \ldots, \hat{H}_K]$ where $\hat{H} = B_K H_K \in \mathbb{C}^{M_R \times N_U}$ represents the composite channel of the $k^{th}user$. Finally, $C_K = [C_{r1}^T, C_{r2}^T, \ldots, C_{rK}^T]^T$ where $B_{rm} \in \mathbb{C}^{N_U \times KN_U}$ for $m = 1, 2, \ldots, K$ is the receive digital precoder for the m^{th} user-pair, and is designed to cancel the inter-user-interference.

The comparison of the EGT hybrid beamforming (HBF), subspace projection-based AoD BD (SP-AoD-BD), BD-HBF, and conventional BD algorithm with $K = 8, L = 8$, the BS employs $M_T = 16$ RF chains is shown in Fig. 22. Each user employs $M_R = 2$ RF chains to support $D_s = 2$ data streams and the SNR ranges from $-40dB$ to $0dB$. The iterations in SP-AoD-BD are 40. As shown in the simulation, the sum spectral efficiency of the $BD-HBF$ has exceeded the performance of the EGT-HBF and SP-AoD-BD. It indicates that the gap increases with the increase of SNR.

3) FULLY/ PARTIALLY-CONNECTED HYBRID BEAMFORMING

The types of RF architectures in hybrid beamforming are a fully connected structure in which each RF chain is connected to all antenna elements [102] and a partially-connected architecture which connects each RF chain only to a subset of antennas [103] as shown in Fig. 23. The partially connected architecture reduces the hardware complexity with some performance degradation. The fully connected struc-

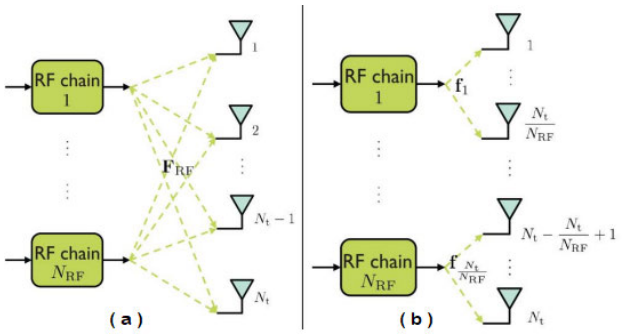

FIGURE 23. Hybrid beamforming (a). Fully connected, (b). Partially connected [102].

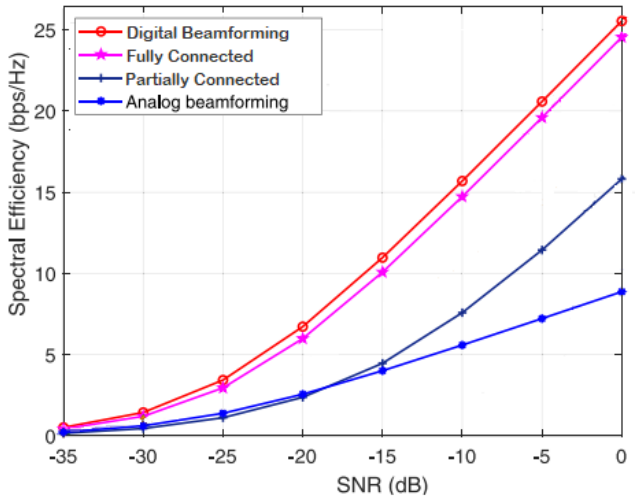

FIGURE 24. Spectral efficiency of different beamforming techniques.

ture performs very close to the fully digital one with many fewer phase shifters (Fig. 24). This makes the fully connected architecture very attractive for practical implementation. The partially-connected architecture, on the other hand, requires lower computational complexity because of hardware implementation simplicity and achieves relatively higher SNR. The performance gap between the partially-connected and fully-connected architecture shrinks as the SNR increases with lower complexity and ease of implementation.

V. CONCLUSION

In this performance evaluation of mmWave-Massive MIMO precoding and beamforming, we have presented comprehensive classifications based on different criteria and scenarios. We survey the most recent works in each of these classifications to include a complete analysis of fundamental ideas and comparative evaluation of mmWave Communication, massive MIMO, and mmWave-massive MIMO systems jointly. It addresses their respective linear precoding, non-linear precoding, and beamforming architectures. The paper focuses on performance and complexity trade-offs and the practical implementation of beamforming techniques for the 5G wireless network.

This article presents an in-depth examination of cutting-edge precoding and beamforming approaches for 5G technologies: massive MIMO, mmWave communications, and mmWave-massive MIMO systems. It also discusses the linear and non-linear precoding techniques used in massive MIMO systems. A comparison of the various linear precoders is also handled. There is also an in-depth argument about non-linear precoders and their performance-complexity profile. Despite being more computationally complex, it is demonstrated that non-linear precoders perform significantly better than linear precoders. On the contrary, as the number of antennas increases and increases, the performance of linear precoders approaches the performance of non-linear precoders with low complexity, cost, and power consumption. With their system models and optimization techniques, millimeter wave-massive MIMO beamforming architectures, such as digital and hybrid analog-digital beamforming, are addressed in detail. In particular, this paper examines a detailed investigation of the potential of hybrid analog/digital beamforming schemes. Besides, it reviews the performance of the fully-connected and partially-connected hybrid analog/digital beamforming architectures. Existing precoding and beamforming techniques are also included with their characteristics, benefits, and drawbacks. The result shows that the partially-connected structure is more efficient in mmWave-massive MIMO communications, which compromises complexity, power consumption, cost, and performance. However, based on the discussions presented in this work, it is evident that there is no single hybrid analog-digital architecture that provides the best trade-off between performance, complexity, and cost. Thus, obtaining the best performance out of the different hybrid beamforming architectures with the lowest cost and complexity requires further study that dynamically designs the architecture according to the channel characteristics and the intended applications.

VI. CHALLENGES AND OPEN ISSUES

Millimeter-wave massive MIMO communications have faced special challenges in the design and performance of hybrid analog-digital beamforming architectures. To achieve the performance of hybrid beamforming close to digital beamforming, the finding of the optimal number of RF chains and analog RF phase shifters, availability of a considerable number of multiplexed data streams and antenna elements is still an existing design problem [21].

On the other hand, in mmWave massive MIMO systems, optimal hybrid analog-digital beamforming gain is achieved only in the case of known channel state information. Direct channel estimation is hindered by fewer RF chains with respect to the number of antenna elements. Furthermore, CSI availability at the transmitter is a challenge. Because the mapping of transmitter and receiver antennas complicates the channel estimation. Thus, obtaining CSI reliably and efficiently while capitalizing on the potential benefits of mmWave massive MIMO is still a major challenge. Some of the issues that have to be addressed are listed as follows.

A. ANTENNA NUMBERS

The number of antennas at both the BS and MS could be much larger than that in conventional MIMO [104]. This makes the CE in mmWave massive MIMO more challenging, even if TDD channel reciprocity is considered. That is, synchronization and calibration errors of RF chains are not trivial to guarantee the channel's reciprocity. Single-antenna users are typically considered in traditional MIMO due to the limited physical size and power capacity. However, both the BS and MS may be equipped with many antennas in mmWave massive MIMO. Thus, precoding in the uplink and combining in the downlink are also important for link reliability. That is, CSI known by both the BS and MS is required, which indicates another challenge that CSI acquired in the uplink by leveraging the channel reciprocity should also be fed back to the MS [41].

B. HARDWARE CONSTRAINTS

The hardware cost and energy consumption of transceivers (high speed ADCs/DAC converters, synthesizers, mixers) in mmWave communications are much higher than those in traditional MIMO systems [19]. Hence, massive low-cost antennas with a limited number of expensive RF chains can be an appealing transceiver structure for mmWave massive MIMO.

C. LOW SNR

In mmWave communications, the bandwidth could be hundreds of MHz or even multiple GHz, which introduces much more thermal noise. In addition, the strong signal directivity of mmWave makes the CE difficult due to the low SNR before beamforming which needs further detailed research for practical implementation [41].

D. MULTI-USER COMMUNICATION

The application of mmWave communications to wireless systems implies considering multi-user communication. Most investigations either consider a single user or several users with a single antenna per user, which is not implied for practical implementation. In addition, further research that considers practical realization must be taken to address the pilot contamination problem that might limit the number of scheduled users and degrade the efficiency of the channel estimation.

REFERENCES

[1] Y. Yang, J. Xu, G. Shi, and C.-X. Wang, *5G Wireless Systems Simulation and Evaluation Techniques*. Ottawa, ON, Canada: Springer, 2018.
[2] P. SinghParihar, R. Saraswat, and S. Maheshwari, "Energy and spectral efficiency of very large multiuser MIMO systems," *Int. J. Comput. Appl.*, vol. 111, no. 5, pp. 4–7, Feb. 2015.
[3] P. S. Srivastav, L. Chen, and A. H. Wahla, "On the performance of efficient channel estimation strategies for hybrid millimeter wave MIMO system," *Entropy*, vol. 22, no. 10, p. 1121, Oct. 2020, doi: 10.3390/e22101121.
[4] J. Huang, R. Feng, J. Sun, C.-X. Wang, W. Zhang, and Y. Yang, "Multi-frequency millimeter wave massive MIMO channel measurements and analysis," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.
[5] T. E. Bogale and L. B. Le, "Massive MIMO and mmWave for 5G wireless HetNet: Potential benefits and challenges," *IEEE Veh. Technol. Mag.*, vol. 11, no. 1, pp. 64–75, Mar. 2016.
[6] J. Hoydis, S. T. Brink, and M. Debbah, "Massive MIMO: How many antennas do we need?" in *Proc. 49th Annu. Allerton Conf. Commun., Control, Comput.*, 2011, pp. 545–550.
[7] T. Kebede, A. Kassaw, Y. Wondie, and J. Stenbrunn, "Joint evaluation of spectral efficiency, energy efficiency and transmission reliability in massive MIMO systems," in *Proc. 7th EAI Int. Conf. Adv. Sci. Technol.*, Bahir Dar, Ethiopia, 2020, pp. 424–435.
[8] T. Parfait, Y. Kuang, and K. Jerry, "Performance analysis and comparison of ZF and MRT based downlink massive MIMO systems," in *Proc. 6th Int. Conf. Ubiquitous Future Netw. (ICUFN)*, Jul. 2014, pp. 383–388.
[9] N. Fatema, G. Hua, Y. Xiang, D. Peng, and I. Natgunanathan, "Massive MIMO linear precoding: A survey," *IEEE Syst. J.*, vol. 12, no. 4, pp. 3920–3931, Dec. 2018.
[10] M. A. B. Mohammad, A. A. Osman, and N. A. A. Elhag, "Performance comparison of MRT and ZF for single cell downlink massive MIMO system," in *Proc. Int. Conf. Comput., Control, Netw., Electron. Embedded Syst. Eng. (ICCNEEE)*, Sep. 2015, p. 52–56.
[11] L. Liang, W. Xu, and X. Dong, "Low-complexity hybrid precoding in massive multiuser MIMO systems," *IEEE Wireless Commun.*, vol. 3, no. 6, pp. 653–656, Dec. 2014.
[12] H. Q. Ngo, "Massive MIMO: Fundamentals and system designs," Linkoping Univ., Linköping, Sweden, Ph.D. dissertation, Linkoping Stud. Sci. Technol., Tech. Rep., 2015.
[13] L. Zhang, L. Gui, K. Ying, and Q. Qin, "Clustering based hybrid precoding design for multi-user massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 68, no. 12, pp. 12164–12178, Dec. 2019.
[14] S. Hur, T. Kim, D. J. Love, J. V. Krogmeier, T. A. Thomas, and A. Ghosh, "Millimeter wave beamforming for wireless backhaul and access in small cell networks," *IEEE Trans. Commun.*, vol. 61, no. 10, pp. 4391–4403, Oct. 2013.
[15] Z. Kuai, T. Wang, and S. Wang, "Transmit antenna selection in massive MIMO systems: An online learning framework," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Aug. 2019, pp. 496–501.
[16] F. Jameel, M. Haider, and A. Butt, "Massive MIMO: A survey of recent advances, research issues and future directions," in *Proc. Int. Symp. Recent Adv. Elect. Eng. (RAEE)*, Oct. 2017, pp. 1–6.
[17] A. Shaikh and M. J. Kaur, "Comprehensive survey of massive MIMO for 5G communications," in *Proc. Adv. Sci. Eng. Technol. Int. Conf. (ASET)*, Mar. 2019, pp. 1–5.
[18] E. Bjornson, J. Hoydis, and L. Sanguinetti, "Massive MIMO has unlimited capacity," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 574–590, Jan. 2018.
[19] S. A. Busari, K. M. S. Huq, S. Mumtaz, L. Dai, and J. Rodriguez, "Millimeter-wave massive MIMO communication for future wireless systems: A survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 2, pp. 836–869, 2nd Quart., 2018.
[20] A. Acheampong and N. Martey, "A comprehensive study of optimal linear pre-coding schemes for a massive MU-MIMO downlink system: A survey," *Int. J. Comput.*, vol. 32, no. 1, pp. 21–33, 2019.
[21] A. N. Uwaechia and N. M. Mahyuddin, "A comprehensive survey on millimeter wave communications for fifth-generation wireless networks: Feasibility and challenges," *IEEE Access*, vol. 8, pp. 62367–62414, 2020.
[22] A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, and S. L. H. Nguyen, "Hybrid beamforming for massive MIMO: A survey," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 134–141, Sep. 2017.
[23] M. Soleimani, R. C. Elliott, W. A. Krzymien, J. Melzer, and P. Mousavi, "Hybrid beamforming for mmWave massive MIMO systems employing DFT-assisted user clustering," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 11646–11658, Oct. 2020.
[24] M. A. Albreem, A. H. Al Habbash, A. M. Abu-Hudrouss, and S. S. Ikki, "Overview of precoding techniques for massive MIMO," *IEEE Access*, vol. 9, pp. 60764–60801, 2021.
[25] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1617–1655, 3rd Quart., 2016.
[26] R. Vannithamby and S. Talwar, "Towards 5G applications," in *Requirements and Candidate Technologies*. Hoboken, NJ, USA: Wiley, 2017.
[27] K. Hassan, M. Masarra, M. Zwingelstein, and I. Dayoub, "Channel estimation techniques for millimeter-wave communication systems: Achievements and challenges," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 1336–1363, 2020.

[28] M. A. Albreem, M. Juntti, and S. Shahabuddin, "Massive MIMO detection techniques: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3109–3132, 4th Quart., 2019.
[29] E. Bjornson, E. G. Larsson, and M. Debbah, "Massive MIMO for maximal spectral efficiency: How many users and pilots should be allocated?" *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1293–1308, Feb. 2016.
[30] E. Björnson, E. G. Larsson, and T. L. Marzetta, "Massive MIMO: Ten myths and one critical question," *IEEE Commun. Mag.*, vol. 54, no. 2, pp. 114–123, Feb. 2016.
[31] G. Erik Larsson and L. V. D. Perre, "Massive MIMO for 5G," *IEEE 5G Tech Focus*, vol. 1, no. 1, p. 1, Mar. 2017.
[32] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.
[33] X. Gao, O. Edfors, F. Rusek, and F. Tufvesson, "Linear pre-coding performance in measured very-large MIMO channels," in *Proc. IEEE Veh. Technol. Conf. (VTC Fall)*, Sep. 2011, pp. 1–5.
[34] J. Wu, "Research on massive mimo key technology in 5G," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 466, no. 1, 2018, Art. no. 012083, doi: 10.1088/1757-899X/466/1/012083.
[35] O. Elijah, C. Y. Leow, T. A. Rahman, S. Nunoo, and S. Z. Iliya, "A comprehensive survey of pilot contamination in massive MIMO—5G system," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 905–923, 2nd Quart., 2016.
[36] A. Zaib, M. Masood, A. Ali, W. Xu, and T. Y. Al-Naffouri, "Distributed channel estimation and pilot contamination analysis for massive MIMO-OFDM systems," *IEEE Trans. Commun.*, vol. 64, no. 11, pp. 4607–4621, Nov. 2016.
[37] L. Varna Babu, L. Mathews, and S. Sakuntala Pillai, "Performance analysis of linear and nonlinear precoding in MIMO systems," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 4, no. 6, pp. 373–376, Jun. 2015.
[38] P. C. Buelga and R. B. Nielsen, "Near field channel modelling through large aperture massive array systems and low-complexity preprocessing robust to local mobility," M.S. thesis, Dept. Electron. Syst., Aalborg Univ., Aalborg, Denmark, 2014.
[39] J. Brady and A. Sayeed, "Beamspace MU-MIMO for high-density gigabit small cell access at millimeter-wave frequencies," in *Proc. IEEE 15th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jun. 2014, pp. 80–84.
[40] D. Darsena, G. Gelli, and F. Verde, *Beamforming and Precoding Techniques*. Hoboken, NJ, USA: Wiley, 2020, doi: 10.1002/9781119471509.w5GRef020.
[41] S. Mumtaz, J. Rodriguez, and L. Dai, *MmWave Massive MIMO: A Paradigm for 5G*. Amsterdam, The Netherlands: Elsevier, 2017.
[42] A. M'uller, A. Kammoun, E. Bjornson, and M. Debbah, "Efficient linear precoding for massive MIMO systems using truncated polynomial expansion," in *Proc. IEEE 8th Sensor Array Multichannel Signal Process. Workshop (SAM)*, Jun. 2014, pp. 273–276.
[43] R. Mai, "Hybrid RF-baseband precoding/combining for massive MIMO wireless communications," Ph.D. dissertation, McGill Univ., Montreal, QC, Canada, 2018.
[44] X. Gao, O. Edfors, F. Rusek, and F. Tufvesson, "Linear pre-coding performance in measured very-large MIMO channels," in *Proc. IEEE Veh. Technol. Conf. (VTC Fall)*, Sep. 2011, pp. 1–5.
[45] S. Sobana and K. Meena Alias Jeyanthi, "Linear and non linear precoding techniques for MIMO broadcast channel," *J. Eng. Appl. Sci.*, vol. 9, no. 7, pp. 273–282, Oct. 2014.
[46] C. Shepard, H. Yu, N. Anand, L. E. Li, T. Marzetta, R. Yang, and L. Zhong, "Practical many antenna basestations," in *Proc. 18th Annu. Int. Conf. Mobile Comput. Netw. (MobiCom)*, 2012, pp. 53–64,
[47] W. Tan, W. Huang, X. Yang, Z. Shi, W. Liu, and L. Fan, "Multiuser precoding scheme and achievable rate analysis for massive MIMO system," *EURASIP J. Wireless Commun. Netw.*, vol. 2018, no. 1, pp. 1–12, Dec. 2018.
[48] Q.-U.-A. Nadeem, A. Kammoun, M. Debbah, and M.-S. Alouini, "Asymptotic analysis of regularized zero-forcing in double scattering channels," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–7.
[49] A. Kammoun, A. Müller, E. Bjornson, and M. Debbah, "Linear precoding based on polynomial expansion: Large-scale multi-cell MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 861–875, Oct. 2014.
[50] A. Mueller, A. Kammoun, E. Björnson, and M. Debbah, "Linear precoding based on polynomial expansion: Reducing complexity in massive MIMO," *EURASIP J. Wireless Commun. Netw.*, vol. 2016, no. 1, Dec. 2016, Art. no. 63, doi: 10.1186/s13638-016-0546-z.
[51] C. Padmaja and B. L. Malleswari, "Bit error rate analysis of 4G communication systems," in *Proc. 13th Int. Conf. Wireless Opt. Commun. Netw. (WOCN)*, Jul. 2016, pp. 1–5.
[52] H. B. Kassa, T. K. Engda, and E. Y. Menta, "Evaluation of spectral vs energy efficiency tradeoff considering transmission reliability in cellular networks," in *Proc. Int. Telemetering Conf.*, vol. 52, Arizona, CA, USA, 2016, pp. 1–5.
[53] A. Hilario-Tacuri and A. Tamo, "BER performance of mm-Wave based systems in rainfall scenarios," in *Proc. Int. Conf. Electron., Electr. Eng. Comput. (INTERCON)*, Aug. 2018, pp. 1–4.
[54] Y. Zhang, Q. Cui, and N. Wang, "Energy efficiency maximization for CoMP joint transmission with non-ideal power amplifiers," in *Proc. IEEE 28th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Oct. 2017, pp. 1–6.
[55] L. Zhao, K. Zheng, H. Long, and H. Zhao, "Performance analysis for downlink massive MIMO system with ZF precoding," *Trans. Emerg. Telecommun. Technol.*, vol. 25, no. 12, pp. 1219–1230, Dec. 2014.
[56] A. Saravanan, "Error rate performance comparisons of precoding schemes over fading channels for multiuser MIMO," *Int. J. Electron. Commun. Eng.*, vol. 11, no. 4, pp. 491–494, 2017.
[57] V. P. Selvan, M. S. Iqbal, and H. S. Al-Raweshidy, "Performance analysis of linear precoding schemes for very large multi-user MIMO downlink system," in *Proc. 4th, Ed., Int. Conf. Innov. Comput. Technol. (INTECH)*, Aug. 2014, pp. 219–224.
[58] T. Marzetta, "Performance evaluation of linear precoding techniques for downlink massive MIMO," *IEEE Sysyt. J.*, vol. 8, no. 3, pp. 1219–1230, Jan. 2014.
[59] Y. Huang, S. He, J. Wang, and J. Zhu, "Spectral and energy efficiency tradeoff for massive MIMO," *IEEE Trans. Veh. Technol.*, vol. 67, no. 8, pp. 6991–7002, Aug. 2018.
[60] J. Tang, D. K. C. So, E. Alsusa, and K. A. Hamdi, "Resource efficiency: A new paradigm on energy efficiency and spectral efficiency tradeoff," *IEEE Trans. Wireless Commun.*, vol. 13, no. 8, pp. 4656–4669, Aug. 2014.
[61] M. Wu, B. Yin, G. Wang, C. Dick, J. R. Cavallaro, and C. Studer, "Large-scale MIMO detection for 3GPP LTE: Algorithms and FPGA implementations," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 916–929, Oct. 2014.
[62] S. Begashaw, X. Shao, E. Visotsky, F. Vook, and A. Ghosh, "Evaluation of tomlinson-harashima precoding for 5G massive MU-MIMO," in *Proc. IEEE 5G World Forum (5GWF)*, Jul. 2018, pp. 77–82.
[63] B. Saradka, S. Bhashyam, and A. Thangaraj, "A dirty paper coding scheme for the multiple input multiple output broadcast channel," in *Proc. Nat. Conf. Commun. (NCC)*, Feb. 2012, pp. 1–5.
[64] S. Jacobsson, G. Durisi, M. Coldrey, T. Goldstein, and C. Studer, "Quantized precoding for massive MU-MIMO," *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 4670–4684, Nov. 2017.
[65] R. D. Wesel and J. M. Cioffi, "Achievable rates for tomlinson-harashima precoding," *IEEE Trans. Inf. Theory*, vol. 44, no. 2, pp. 824–831, Mar. 1998.
[66] P. Thakor and R. Sathvara, "Performance of Tomlinson-Harashima precoding and dirty paper coding for broadcast channels in MU-MIMO," *Int. Res. J. Eng. Techno.*, vol. 3, no. 4, pp. 2458–2462, 2016.
[67] Y. Chen, "Low complexity precoding schemes for massive MIMO systems," Ph.D. dissertation, School Eng., Newcastle Univ., Newcastle Upon Tyne, U.K., 2019.
[68] M. Huang, X. Zhang, S. Zhou, and J. Wang, "Tomlinson-harashima precoding in multiuser MIMO systems with imperfect channel state information," in *Proc. IEEE Global Telecommun. Conf.*, Nov. 2007, pp. 2806–2810.
[69] A. R. Flores, B. Clerckx, and R. C. de Lamare, "Tomlinson-harashima precoded rate-splitting for multiuser multiple-antenna systems," in *Proc. 15th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Aug. 2018, pp. 1–6.
[70] L. Du, L. Li, P. Zhang, D. Miao, and Z. Liu, "Vector perturbation precoding under imperfect CSI and inaccurate power scaling factors," *IEEE Access*, vol. 7, pp. 89162–89171, 2019.
[71] A.-A. Lu, X. Gao, W. Zhong, C. Xiao, and X. Meng, "Robust transmission for massive MIMO downlink with imperfect CSI," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5362–5376, Aug. 2019.
[72] S. P. Herath, D. H. N. Nguyen, and T. Le-Ngoc, "Vector perturbation precoding for multi-user CoMP downlink transmission," *IEEE Access*, vol. 3, pp. 1491–1502, 2015.
[73] C. Masouros, M. Sellathurai, and T. Ratnarajah, "Low complexity vector precoding for fast fading MIMO downlinks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2013, pp. 3265–3269.

[74] A. Agnihotri and B. Gupta, "Performance evaluation of linear/non-linear precoding schemes for downlink multi-user MIMO systems," in *Proc. Int. Conf. Ind. Instrum. Control (ICIC)*, May 2015.
[75] Q. C. Li, H. Niu, A. T. Papathanassiou, and G. Wu, "5G network capacity: Key elements and technologies," *IEEE Veh. Technol. Mag.*, vol. 9, no. 1, pp. 71–78, Mar. 2014.
[76] X. Wang, L. Kong, F. Kong, F. Qiu, M. Xia, S. Arnon, and G. Chen, "Millimeter wave communication: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 1616–1653, 3rd Quart., 2018.
[77] A. Das and S. Kolangiammal, "Performance analysis of millimeter wave communication system using 256-QAM and 512-QAM techniques," in *Proc. Int. Conf. Commun. Signal Process. (ICCSP)*, Apr. 2017, pp. 360–364.
[78] M. Xiao, S. Mumtaz, Y. Huang, L. Dai, Y. Li, M. Matthaiou, G. K. Karagiannidis, E. Björnson, K. Yang, C.-L. I, and A. Ghosh, "Millimeter wave communications for future mobile networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1909–1935, Sep. 2017.
[79] A. Ghosh, T. A. Thomas, M. C. Cudak, R. Ratasuk, P. Moorut, F. W. Vook, T. S. Rappaport, G. R. MacCartney, S. Sun, and S. Nie, "Millimeter-wave enhanced local area systems: A high-data-rate approach for future wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1152–1163, Jun. 2014.
[80] D.-W. Yue, S. Xu, and H. H. Nguyen, "Diversity gain of millimeter-wave massive MIMO systems with distributed antenna arrays," *EURASIP J. Wireless Commun. Netw.*, vol. 2019, no. 1, pp. 1–13, Dec. 2019.
[81] R. Yahia Ramadan and H. Minn, "Novel hybrid precoding designs for mmWave multiusern systems with partial channel knowledge," in *Proc. IEEE Global Commun. Conf.*, Dec. 2017, pp. 1–6.
[82] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, Jr., "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 831–846, Oct. 2014.
[83] Y. Huang, C. Liu, Y. Song, and X. Yu, "DFT codebook-based hybrid precoding for multiuser mmWave massive MIMO systems," *EURASIP J. Adv. Signal Process.*, vol. 2020, no. 1, pp. 1–13, Dec. 2020, doi: 10.1186/s13634-020-00669-4.
[84] A. L. Swindlehurst, E. Ayanoglu, P. Heydari, and F. Capolino, "Millimeter-wave massive MIMO: The next wireless revolution?" *IEEE Commun. Mag.*, vol. 52, no. 9, pp. 56–62, Sep. 2014.
[85] Y. Sun and C. Qi, "Weighted sum-rate maximization for analog beamforming and combining in millimeter wave massive MIMO communications," *IEEE Commun. Lett.*, vol. 21, no. 8, pp. 1883–1886, Aug. 2017.
[86] G. Hegde, "Energy-efficient and robust hybrid analog-digital precoding for massive MIMO systems," Ph.D. dissertation, Dept. Elect. Eng. Inf. Technol., Darmstadt Tech. Univ., Darmstadt, Germany, 2019. [Online]. Available: https://tuprints.ulb.tu-darmstadt.de/9210/
[87] Y. Song, S. Wu, W. Zhang, and H. Zhang, *Millimeter Wave Massive MIMO Channel Estimation and Tracking*. Cham, Switzerland: Springer, 2019.
[88] I. Ahmed, H. Khammari, A. Shahid, A. Musa, K. S. Kim, E. De Poorter, and I. Moerman, "A survey on hybrid beamforming techniques in 5G: Architecture and system model perspectives," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 3060–3097, 4th Quart., 2018.
[89] X. Gao, L. Dai, S. Han, C.-L. I, and R. W. Heath, Jr., "Energy-efficient hybrid analog and digital precoding for mmWave MIMO systems with large antenna arrays," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 998–1009, Apr. 2016.
[90] X. Lu, W. Yang, Y. Cai, and X. Guan, "Comparison of CS-based channel estimation for millimeter wave massive MIMO systems," *Appl. Sci.*, vol. 9, no. 20, p. 4346, Oct. 2019, doi: 10.3390/app9204346.
[91] F. Sohrabi and W. Yu, "Hybrid digital and analog beamforming design for large-scale antenna arrays," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 501–513, Apr. 2016.
[92] A. N. Uwaechia, N. M. Mahyuddin, M. F. Ain, N. M. A. Latiff, and N. F. Za'bah, "On the spectral-efficiency of low-complexity and resolution hybrid precoding and combining transceivers for mmWave MIMO systems," *IEEE Access*, vol. 7, pp. 109259–109277, 2019.
[93] S. Han, C.-L. I, Z. Xu, and C. Rowell, "Large-scale antenna systems with hybrid analog and digital beamforming for millimeter wave 5G," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 186–194, Jan. 2015.
[94] R. Mendez-Rial, C. Rusu, N. Gonzalez-Prelcic, A. Alkhateeb, and R. W. Heath, Jr., "Hybrid MIMO architectures for millimeter wave communications: Phase shifters or switches?" *IEEE Access*, vol. 4, pp. 247–267, 2016.
[95] T. Kebede, Y. Wondie, and J. Steinbrunn, "Performance evaluation of MillimeterWave-massive MIMO with beamforming techniques," in *Proc. Int. Symp. Netw., Comput. Commun. (ISNCC)*, Oct. 2021, pp. 1–8.
[96] Y. Liu, Q. Feng, Q. Wu, Y. Zhang, M. Jin, and T. Qiu, "Energy-efficient hybrid precoding with low complexity for mmwave massive MIMO system," *IEEE Access*, vol. 7, pp. 95021–95032, 2019, doi: 10.1109/ACCESS.2019.2928559.
[97] A. M. Hamed and R. K. Rao, "Spectral and energy efficiencies in mmWave cellular networks for optimal utilization," *Wireless Commun. Mobile Comput.*, vol. 2018, pp. 1–11, Oct. 2018, doi: 10.1155/2018/3097094.
[98] J. Chen, H. Chen, H. Zhang, and F. Zhao, "Spectral-energy efficiency tradeoff in relay-aided massive MIMO cellular networks with pilot contamination," *IEEE Access*, vol. 4, pp. 5234–5242, 2016.
[99] Q. Zhang, Y. Liu, G. Xie, J. Gao, and K. Liu, "An efficient hybrid diagonalization for multiuser mmWave massive MIMO systems," in *Proc. 11th Global Symp. Millim. Waves (GSMM)*, May 2018, pp. 1–6.
[100] W. Ni and X. Dong, "Hybrid block diagonalization for massive multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 64, no. 1, pp. 201–211, Jan. 2016.
[101] I. Altoobchi and M. A. Mangoud, "Investigations of beamforming designs and millimeter wave channel modeling for multiuser MIMO systems," in *Proc. 9th IEEE-GCC Conf. Exhib. (GCCCE)*, May 2017, pp. 1–5.
[102] J. Zhang, X. Yu, and K. B. Letaief, "Hybrid beamforming for 5G and beyond millimeter-wave systems: A holistic view," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 77–91, 2020.
[103] O. E. Ayach, R. W. Heath, Jr., S. Abu-Surra, S. Rajagopal, and Z. Pi, "Low complexity precoding for large millimeter wave MIMO systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2012, pp. 3724–3729.
[104] A. Liao and Z. Gao, "Super-resolution channel estimation for mmWave massive MIMO," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–5.
[105] K. Tsuge, Y. Chang, K. Fukawa, S. Suyama, and Y. Okumura, "Parameter estimation for block diagonalization based hybrid beamforming in massive MIMO communications," in *Proc. IEEE 90th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2019, pp. 1–5.
[106] E. G. Larsson and E. Bjornson. Accessed: May 24, 2018. [Online]. Available: https://ma-mimo.ellintech.se/2017/10/03/what-is-the-difference-between-beamforming-and-precoding/
[107] Q. Zhang, S. Jin, M. McKay, D. Morales-Jimenez, and H. Zhu, "Power allocation schemes for multicell massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 5941–5955, Nov. 2015.
[108] A. Li and C. Masouros, "Hybrid precoding and combining design for millimeter-wave multi-user MIMO based on SVD," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–7.
[109] T. Kebede, Y. Wondie, and J. Steinbrunn, "Massive MIMO linear precoding techniques performance assessment," in *Proc. Int. Symp. Netw., Comput. Commun. (ISNCC)*, Oct. 2021, pp. 1–8.

TEWELGN KEBEDE was born in Echegie Gelila Kebele, East Gojjam, Amhara, Ethiopia, in 1974. He received the B.Sc. degree in electrical-electronics technology from Adama University, in 2004, the B.Sc. degree in computer science from the Infolink University College, in 2010, and the M.Sc. degree in communication engineering and networking from Hawassa University, in 2015. He is currently pursuing the joint Ph.D. degree in communication engineering with the Addis Ababa Institute of Technology, Addis Ababa University, Ethiopia, and the Kempten University of Applied Sciences, Germany.

Since 2011, he has been a Lecturer with the Department of Electrical and Computer Engineering, Hawassa University. From 2008 to 2011, he was a United Nations Volunteer ICT Specialist. He is the author of about five articles. His research interests include massive MIMO and mmWave communications for 5G wireless networks.

YIHENEW WONDIE was born in Eastern Gojjam, Amhara, Ethiopia, in 1980. He received the B.Sc. degree in electrical engineering from Addis Ababa University, in 2003, the M.Sc. degree in telecommunications engineering from the Graduate School of Telecommunications and Information Technology, in 2009, and the Ph.D. degree in electrical engineering and computer science specializing in communication engineering from Kumamoto University, Japan, in 2015.

Since 2015, he has been working with the Addis Ababa Institute of Technology (AAiT), as an Assistant Professor. He has authored more than 25 articles. His research interests include almost all the advances in wireless communications, such as massive MIMO, mmWave communications, cooperative wireless communications, distributed antenna systems, and precoding and multiplexing techniques in 5G and beyond. In 2014, he won the Excellent Presentation Award of the IEEE Fukuoka Section, Japan.

JOHANNES STEINBRUNN was born in Albersdorf, Vilshofen, Germany, in October 1945. He received the M.Sc. (Dipl.-Ing.) in electrical engineering communication and automation engineering from the Karlsruhe Institute of Technology (KIT), Germany, in September 1972, the Ph.D. (Dr.-Ing.) degree from the University of Erlangen, Germany, in November 1978, and the Honorary doctorate (Dr.H.C.) degrees from the Technical University of Tallinn, Estonia, in 2003, and the University of Ulster, U.K., in 2008. Afterward, in 1979, he was employed with the University of Georgia, Athens, USA, as an Exchange Scientist in the field of computer science. Back in Germany, he worked in the aerospace industry in the research and development of missile technology control. In 1982, the Kempten University of Applied Sciences, Germany, appointed him as a Professor in the field of electrical drive and automation engineering. He conducted long-term research and teaching in different European and Asian countries. He was also the Dean of the Faculty of Electrical Engineering and the Vice-President of the Kempten University of Applied Sciences. From 2011 to 2013, he developed the Institute of Technology (IoT) as the Scientific Director of Hawassa University, Ethiopia.

HAILU BELAY KASSA received the B.Sc. degree in electrical engineering from Defense University, Ethiopia, in 2005, the M.S. degree in telecommunications engineering from Addis Ababa University, Addis Ababa, Ethiopia, in 2009, and the Dr. Eng. degree in electrical engineering from Morgan State University, Baltimore, MD, USA, in 2018. He serves or has served on the review of several wireless communication systems related IEEE international conferences and journals. He also worked in different industries as a Software Engineer, such as at Dallol Group Inc., Atlanta, USA, in 2016, and Unicom Solution Plc., Ethiopia, in 2014 and 2015. He is currently working as a Researcher with Morgan State University.

KEVIN T. KORNEGAY is currently a Professor in electrical and computer engineering with Morgan State University. At the time of his MIT appointment, he was an Assistant Professor with Purdue University. His research interests include big band gap semiconductor devices, smart power electronics, and power electronic building blocks (PEBBs), wireless MEMS and integrated electronics for harsh environments, and VLSI design and CAD for VLSI, radio frequency and millimeter-wave integrated circuit design, high-speed circuits, broadband wired and wireless communication systems, and cyber-physical systems.