A design framework for all-digital mmWave massive MIMO with per-antenna nonlinearities

Mohammed Abdelghany, Ali A. Farid, Maryam Eslami Rasekh, Upamanyu Madhow, *Fellow, IEEE*, and Mark J. W. Rodwell, *Fellow, IEEE*.

***Abstract*—Millimeter wave MIMO combines the benefits of compact antenna arrays with a large number of elements and massive bandwidths, so that fully digital beamforming has the potential of supporting a large number of simultaneous users with *per user* data rates of multiple gigabits/sec (Gbps). In this paper, we develop an analytical model for the impact of nonlinearities in such a system, and illustrate its utility in providing hardware design guidelines regarding two key challenges: the low available precision of analog-to-digital conversion at high sampling rates, and nonlinearities in ultra-high speed radio frequency (RF) and baseband circuits. We consider linear minimum mean square error (LMMSE) reception for a multiuser MIMO uplink, and provide performance guarantees based on two key concepts: (a) summarization of the impact of per-antenna nonlinearities via a quantity that we term the "intrinsic SNR", (b) using linear MMSE performance in an ideal system without nonlinearities to bound that in our non-ideal system. For our numerical results, we employ nominal parameters corresponding to outdoor picocells operating at a carrier frequency of 140 GHz, with a data rate of 10 Gbps per user.**

***Keywords*—All-digital massive MIMO uplink design, LoS channel, Nonlinearity (P_{1dB}), Low-precision ADC, Load factor, LMMSE.**

I. Introduction

The potential of massive scale multiuser MIMO for meeting the ever-increasing demand for wireless mobile data is well understood [1], [2]. Massive MIMO becomes particularly attractive as we move up in the frequency spectrum toward millimeter wave (mmWave) and terahertz (THz) frequencies, where bandwidth is plentiful, and the wavelength is small enough to fit a large number of antennas on moderately sized platforms. By utilizing these advantages, mmWave massive MIMO can potentially support tens or hundreds of simultaneous users with per-user data rates of multiple gigabits/second (Gbps).

Two key bottlenecks to realizing this potential are the cost and power consumption of radio frequency (RF) frontends at high carrier frequencies and analog-to-digital conversion at large bandwidths. Due to the higher power consumption of high-speed RF chains, mmWave prototypes have thus far opted against fully digital arrays [3], and much of the recent research and development has focused on analog RF beamforming [4]–[9], supporting a single user at a time, or hybrid beamforming [10]–[15], where the number of supported users equals the number of RF chains, typically set to be much smaller than the number of antennas. However, recent advances in mmWave silicon hardware imply that scaling the number of RF chains with the number of antennas is on the cusp of feasibility, opening up the possibility of digital beamforming, where the number of supported users scales linearly with the number of antennas. By reducing dynamic range requirements and increasing amplifier loading, we can boost the power efficiency of each RF chain enough to allow scaling to fully digital arrays with hundreds of elements, but at the cost of increasing nonlinearity in the RF chain. Similarly, drastically reduced precision can be used to control the cost and power consumption of analog-to-digital conversion, as well as that of communication and computation on the digital backend. Robust system design in the presence of such nonlinearities therefore plays a critical role in scaling digital beamforming to mmWave massive MIMO. Our goal in this paper is to provide an analytical framework for quantifying the *system-level* impact of such nonlinearities, and to demonstrate how the increased degrees of freedom help relax linearity requirements and hardware specifications.

We consider, as a running example, a 256-element linear array at 140 GHz carrier frequency, with a symbol rate of 5 Gbaud with QPSK modulation for each supported user. For a load factor (defined as the ratio of the number of simultaneous users to the number of antennas) ranging from $\frac{1}{16}$ (16 users) to $\frac{1}{2}$ (128 users), the aggregate data rate ranges from 160 Gbps to 1.28 Tbps! However, scaling to this regime is challenging: wideband RF and baseband circuits scaled via relatively low-end silicon semiconductor processes (e.g., CMOS) exhibit significant nonlinearities, while the analog-to-digital converters (ADCs) available at multi-GHz sampling rates have relatively low precision. We provide in this paper an analytical framework that enables designers to determine the permissible levels of such nonlinearities for their desired system-level performance guarantees. As we shall show, the RF linearity and ADC precision requirements for load factor $\frac{1}{2}$ are quite stringent, while reducing the load factor to $\frac{1}{16}$ results in significantly relaxed hardware specifications. For a given number of users to be supported, therefore, a massive MIMO architecture can be leveraged to overcome severe nonlinearities, by increasing the number of antenna elements in order to reduce the load factor. We note that, besides the enormous aggregate throughput from supporting multiple simultaneous users, recent work also indicates that an all-digital solution can be more efficient in terms of hardware power consumption and

M. Abdelghany A. Farid, M.E. Rasekh, U. Madhow, and M. Rodwell are with the Department of Electrical and Computer Engineering, University of California at Santa Barbara, Santa Barbara, CA 93106 USA (e-mail: mabdelghany@ucsb.edu; afarid@ece.ucsb.edu; rasekh@ece.ucsb.edu; madhow@ece.ucsb.edu; rodwell@ucsb.edu).

area compared to a hybrid architecture [16].

A. Contributions

We provide design guidelines based on linear MMSE reception, with an analytical framework based on two core concepts: (a) We use a matched filter bound to show that the impact of per-antenna nonlinearities is effectively summarized by a quantity that we term the *intrinsic SNR,* corresponding to a normalized version of the nonlinearity. Key elements of this characterization are the well-known Bussgang decomposition, an overview of which can be found in [17], and the observation that, even for a moderate number of simultaneous users and without rich scattering, the antenna input is well modeled as zero-mean complex Gaussian. We show that the matched filter bound on the effective SNR for a given user, which captures the effect of the self-noise generated by per-antenna nonlinearities, depends only on four parameters: the user's SNR, the intrinsic SNR, the load factor and a power control factor which summarizes the variations in received signal power across users.
(b) We show that a pessimistic estimate of the degradation in performance due to multiuser interference can be obtained by analyzing (theoretically and/or numerically) an *ideal* system without nonlinearities. This enables us to provide a lower bound on the output signal-to-interference-plus-noise ratio (SINR) of a linear MMSE receiver, accounting for both nonlinearities and multiuser interference.

Combining these two concepts, averaging over the spatial distribution of users, and specializing to an edge user in the cell, allows us to provide analytical guidelines for maximum permissible levels of nonlinearities in order to provide the desired system-level performance guarantees (e.g., on outage probabilities). Consequently, our analysis utilizes the analytical capabilities of the Bussgang decomposition and LMMSE framework to provide a cross-layer design tool that links hardware level specifications to the system level performance metrics of a multi-user massive MIMO system. This enables exploration of fundamental tradeoffs between power consumption and cost of RF frontends and system performance. We consider third order RF and baseband nonlinearities that can be specified using the so-called 1 dB compression point [18], termed P_{1dB}. The per-antenna ADCs for the in-phase and quadrature components are modeled as overloaded uniform quantizers optimized (for a specified number of bits) to minimize the mean square error with zero mean Gaussian input. Using our framework, we are able to provide compact design prescriptions for P_{1dB} and the number of ADC bits. For example, for a load factor of 1/2, the system can work with 4-bit ADC and passband/baseband P_{1dB} of 8.4 dB / 5 dB. On the other hand, 2-bit ADC with passband/baseband P_{1dB} of 1.4 dB / -1 dB suffice to work properly with a load factor of 1/16. We present extensive simulations verifying our analytical predictions and prescriptions.

B. Related Work

While the focus in the present paper is on mmWave massive MIMO, there is a significant body of closely related recent research on the effect of nonlinearities on multiuser massive MIMO at lower carrier frequencies. Most of this prior work also employs Bussgang's theorem [19] to model the effect of nonlinearities, both for uplink reception and downlink precoding. Our discussion here is limited to the literature on uplink massive MIMO, since that is the focus of the present paper, but the design framework for modeling downlink nonlinearities such as power amplifiers and digital-to-analog converters (DACs) is well known to be entirely analogous.

The line of sight (LoS) channel model used in our performance evaluation is different from that in much of this prior work, which employs models that are better matched to the propagation environments at lower frequencies. However, our analytical framework is quite general, and can be used to obtain design prescriptions for lower carrier frequencies as well. Conversely, many of the general observations emerging from prior work at lower carrier frequencies are consistent with the conclusions in the present paper, given a common underlying mathematical framework that employs the Bussgang decomposition and exploits the relaxation of hardware constraints enabled by the increase in the number of antennas. In the following, we briefly review this prior work in order to place the contributions of the present paper in perspective.

The potential for relaxing hardware constraints by increasing the number of antennas is clearly brought out by the theoretical results in [20], which show that the performance degradation due to hardware impairments vanishes asymptotically as the number of base station antennas gets large. The same trend holds for a finite but large number of antennas, as is clear from the results in [21]–[23], which study the spectral efficiency of quantized massive MIMO over frequency nonselective Rayleigh and Rician fading channels using maximum ratio combining. Another interesting conclusion from the simulations of [22] is that, for Rician fading, the system is more vulnerable to drastic quantization as the relative strength of the dominant component increases. Thus, the LoS model considered in this paper may be a worst-case scenario for obtaining design prescriptions regarding nonlinearities.

The impact of imperfect power control for quantized massive MIMO over frequency nonselective channels is included in the analysis in [24], [25]. Using spectral efficiency as a performance measure, an example conclusion from [24] is that 3-bit ADC suffices for a system with 100 antennas serving 10 users at a spectral efficiency of 3.5 bits per channel use, with 4-bit ADCs recommended to handle imperfections in power control and automatic gain control. Similar conclusions are obtained in [25], which shows moderate drops in spectral efficiency due to imperfect power control.

The impact of quantization on multiuser OFDM MIMO over a frequency-selective channel is studied in [26], with a focus on low-complexity channel estimation and data detection. The simulations in this paper show that, for the models considered, 4-bit ADC is sufficient to achieve a near-optimal performance (in terms of packet error rate) for a load factor of 1/8 or lower. More recent work with a similar model [27] employs a Bussgang-based analysis for the joint distortion introduced by nonlinear low-noise amplifiers, phase noise, and finite-resolution ADCs, and demonstrates its accuracy by comparing

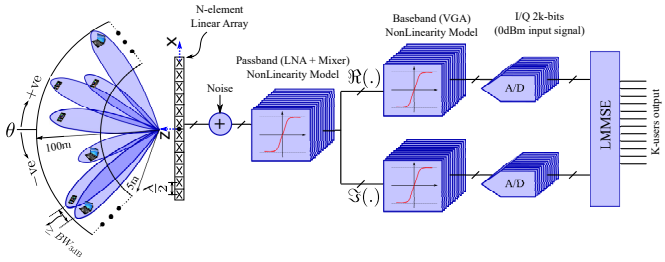


Figure 1: The cell size is constrained radially between $R_{\min}$ and $R_{\max}$ and angularly between $-\pi/3 \leq \theta \leq \pi/3$. $\text{BW}_{3\text{dB}}$ stands for the 3 dB beamwidth. The passband and baseband nonlinearities are modeled by saturated third order polynomials. An overloaded uniform ADC with b bits per dimension, optimized for a zero-mean standard Gaussian random variable, is used. Linear MMSE reception is employed after digitization.

analytical predictions with simulations.

In comparison with the existing literature, the key conceptual novelty in the present paper is that we provide an analytical framework for mapping *system-level* performance goals to *hardware design* prescriptions for per-antenna nonlinearities. The theoretical foundation for this mapping is our observation (Theorem V.2) that an ideal system without nonlinearities provides a means of obtaining pessimistic performance estimates, together with our abstraction of self-noise via intrinsic SNR and the associated matched filter bound (Theorem V.1). Thus, while prior work such as [26], [27] demonstrates the accuracy of Bussgang modeling and assesses design tradeoffs in particular scenarios, we are able to provide a general framework which provides *compact* prescriptions that hardware designers can apply to design RF chains jointly with ADCs, by considering the cascade of passband amplifiers, baseband amplifiers and ADCs as the nonlinearities employed in our performance evaluation. Finally, unlike prior work on fading channels, we employ a LoS model which is a more suitable abstraction for mmWave channels [28]–[31].

A preliminary version of this work has appeared in a conference paper [32]. In this paper, we provide a comprehensive analysis, including proofs that were omitted in [32], along with a more extensive set of numerical results. We also study the impact of power control on our system-level performance objectives. The system model is also different in some details from [32] in order to more closely model the hardware designs that we are currently engaged in: we now include the impact of baseband as well as RF nonlinearities, and consider a more reasonable field of view for the base station array.

II. System Model

Fig. 1 shows the system model. The base station performs horizontal scanning with a 1D half-wavelength spaced N-element array. Let K denotes the number of simultaneous users, and $\beta = \frac{K}{N}$ the *load factor*.

We assume a line-of-sight (LoS) channel between the base station and each mobile. The direction of arrival (DoA) from the k^{th} mobile is denoted by θ_k, and corresponds to spatial frequency $\Omega_k = 2\pi \frac{d_x}{\lambda} \sin\theta_k$, where λ denotes the carrier wavelength and d_x denotes the inter-element spacing, set to $\frac{\lambda}{2}$ in our numerical results. The $N \times 1$ spatial channel for mobile k is given by

$$\mathbf{h}_k = A_k e^{j\phi_k} \left[1\, e^{j\Omega_k}\, e^{j2\Omega_k}\, \ldots\, e^{j(N-1)\Omega_k}\right]^{\intercal}, \tag{1}$$

where ϕ_k is an arbitrary phase shift and $A_k^2 = \left(\frac{\lambda}{4\pi R_k}\right)^2$ depends on the radial location R_k of mobile k, using the Friis formula for path loss.

The cascade of the nonlinearities described in Sections II-B and II-C is modeled as a complex baseband equivalent nonlinearity $g(\cdot)$. The complex baseband received signal vector $\mathbf{z}$ at the base station is therefore given by

$$\mathbf{z} = g(\mathbf{y}) = g\left(\underbrace{\mathbf{H}}_{N\times K}\mathbf{x} + \mathbf{n}\right), \tag{2}$$

where $\mathbf{H} = [\mathbf{h}_1 \mathbf{h}_2 \ldots \mathbf{h}_K]$ is the channel matrix, $\mathbf{x} = [x_1, ..., x_K]^T$ is the vector of symbols (normalized to unit energy: $\mathbb{E}\left[|x_k|^2\right] = 1$) transmitted by the mobiles, $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma_n^2\mathbf{I})$ is the thermal AWGN vector, and $g(\cdot)$ is the effective per-antenna nonlinearity in complex baseband.

Running example: As mentioned in the introduction, for our running example, we set $N = 256$, with load factor β ranging from $\frac{1}{16}$ to $\frac{1}{2}$ (i.e., K ranging from 16 to 128). We provide the link budget analysis for the envisioned system in Appendix A to highlight the feasibility of low-cost silicon hardware realizations.

In the remainder of this section, we characterize the statistics of the received signal at each antenna and describe the nonlinearity models included in our numerical results.

A. Per-antenna Received Signal Statistics

The input to the effective complex baseband nonlinearity $g(\cdot)$ at, say, antenna m, is given by

$$y_m = \sum_{k=1}^{K} A_k e^{j\phi_k} x_k e^{jm\Omega_k}. \tag{3}$$

For a uniform spatial distribution of users over the region of interest, the amplitudes $\{A_k\}$ and spatial frequencies $\{\Omega_k\}$ are independent and identically distributed (i.i.d.). The phases $\{\phi_k\}$ are uniform over $[0, 2\pi]$, and x_k are i.i.d. QPSK symbols. By virtue of the central limit theorem (CLT), the received signal is well modeled as zero-mean complex Gaussian for large K, and jointly Gaussian across antennas. We have verified empirically, via histogram comparisons, quantile-quantile plots and KL divergence computations, that this Gaussian approximation holds for even moderate number of mobiles (e.g., $K = 8$) in all settings that we have considered. Fig. 2 (a) illustrates a comparison between the histogram of the normalized real/imaginary component of the received signal and the standard normal distribution $\mathcal{N}(0, 1)$.

In terms of technical conditions for applying the CLT, we note that it holds for independent, non-identically distributed

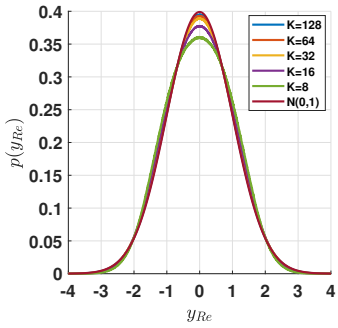


Figure 2: The pdf of the standard normal distribution and the histogram of the normalized real/imaginary part of the received signal at each antenna element when K users transmit.

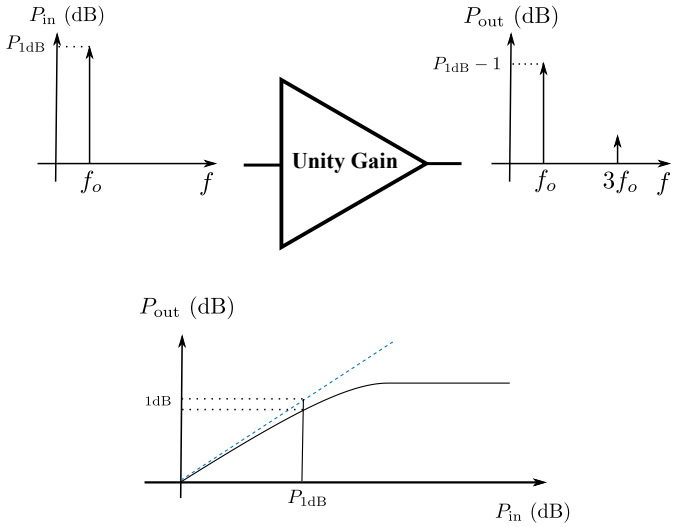
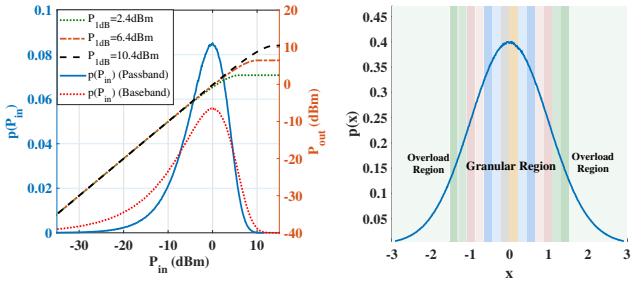


Figure 3: The 1 dB compression point (P_{1dB}) is defined as the input power at which the output power of the desired sinusoid (at f_o) is compressed by 1 dB.

random variables if the variances are bounded [33], which is the case here: with power control, the amplitudes of $\{A_k\}$ are tightly clustered, whereas with no power control they lie within a range of values determined by the maximum and minimum link distance. The randomness in the terms of the sum in (3) results from randomness in channel phases and data modulations, and CLT is applied conditioned on $\{A_k\}$. We average over the realizations of $\{A_k\}$ to determine expected receiver performance and reliability bounds.

B. Passband and Baseband Nonlinearity Model

The passband nonlinearity arises in the low noise amplifier and the mixer, while the baseband nonlinearity is in the variable gain amplifier. We model each nonlinearity as a saturated third-order polynomial function with a nominal gain of unity. The function is parametrized by the 1 dB compression



(a) Third-order nonlinearities (b) Overloaded uniform ADC

Figure 4: (a) Third-order nonlinearities characterized by P_{1dB}, and probability distribution function of instantaneous input power, $\text{p}(\text{P}_{\text{in}})$, for passband and baseband signals. (b) Histogram of I and Q baseband components along with ADC quantization bins.

point (P_{1dB}) [18], defined as the input power of a sinusoid of frequency f_o (taken to be the carrier frequency) at which the output power is reduced by 1 dB relative to the nominal. The concept is illustrated in Fig. 3. A commonly used model for gain saturation using third-order nonlinearity is the cubic soft clipper which can be parametrized by P_{1dB} as follows:

$$g(y(t)) = \begin{cases} y(t)(1 - \frac{0.44|y(t)|^2}{3P_{\text{1dB}}}) & \text{if } |y(t)|^2 \leq \frac{P_{\text{1dB}}}{0.44} \\ \frac{y(t)}{|y(t)|}\sqrt{P_{\text{1dB}}} & \text{if } |y(t)|^2 > \frac{P_{\text{1dB}}}{0.44} \end{cases}. \quad (4)$$

The gain compression for the passband nonlinearity depends on the absolute value of the complex baseband signal, while the gain compression depends on the absolute value of the I and Q components for the baseband nonlinearity. Fig. 4 (a) illustrates the distribution of the input powers of the passband and baseband nonlinearities, along with example input/output (I/O) characteristics. In this work, we consider the nonlinearities to be memoryless and free of phase distortion.

C. ADC Model

After down-conversion, each baseband component is quantized to b bits by an ADC. We design the quantizer to minimize the mean squared error (MSE) assuming that the incoming signal is Gaussian with zero mean and unit variance. An automatic gain control (AGC) precedes the ADC in order to normalize the average power of the input signal to unity, and ensure that the full dynamic range of the ADC is utilized. We employ an overloaded uniform ADC [34]: while the MSE could be improved slightly by designing a non-uniform quantizer for standard Gaussian input, the improvement is slight and has no discernible impact on system-level performance (see Appendix B for a quantitative discussion). Fig. 4 (b) depicts a 4-bit uniform overloaded quantizer.

D. Linear MMSE Detector

We show in a following section that the impact of a per-antenna nonlinearity $g(\cdot)$ can be modeled as additional noise,

leading to an equivalent system model of the form

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \tilde{\mathbf{n}}, \tag{5}$$

where $\tilde{\mathbf{n}}$ is zero mean with variance $(\sigma_n^2 + \nu_g^2)\mathbf{I}$. The value of ν_g is specified in section V. Thus, any adaptive implementation of the linear MMSE receiver automatically accounts for the nonlinearities. The linear MMSE receiver is specified as follows:

$$\hat{\mathbf{x}} = \mathbf{W}\mathbf{y}, \tag{6}$$

where

$$\mathbf{W} = \left(\mathbf{H}^H\mathbf{H} + (\sigma_n^2 + \nu_g^2)\mathbf{I}\right)^{-1}\mathbf{H}^H. \tag{7}$$

The linear MMSE detector has a rich history with well-known properties [35], [36]. In order to provide a self-contained exposition, we state a few properties that are relevant for our present purpose, and sketch their proof in Appendix C.

III. Bussgang Linearization

In order to provide a self-contained exposition, we review Bussgang linearization in the context of our MIMO system.

A. Scalar Bussgang Linearization

For a zero mean complex-valued random variable y and a nonlinearity $g(\cdot)$, a linear MMSE approximation of $g(y)$ by ay satisfies the orthogonality principle [37]:

$$\mathbb{E}((g(y) - ay)y^*) = 0. \tag{8}$$

Standard computations for the linear gain a and the variance of the approximation error $e = g(y) - ay$ yield

$$a = \frac{\mathbb{E}(g(y)y^*)}{\mathbb{E}(|y|^2)}, \tag{9}$$

$$\sigma_g^2 = \mathbb{E}(|e|^2) = \mathbb{E}(|g(y)|^2) - |a|^2\mathbb{E}(|y|^2). \tag{10}$$

Hence, $g(y)$ can be written as

$$g(y) = ay + e, \tag{11}$$

where a and $\mathbb{E}(|e|^2) = \sigma_g^2$ can be computed analytically or empirically for any distribution of y and nonlinear function $g(\cdot)$. Bussgang evaluated a and σ_g^2 for different nonlinear functions when the input y is Gaussian random variable [19]. In our analysis, we consider the function $g(\cdot)$ described in (4) for the overall RF chain nonlinearity.

B. Vector Bussgang Linearization

The main part of Bussgang's theorem in [19], and its extension to the complex domain in [38], is the preservation of covariance structure under nonlinearities for jointly Gaussian random variables:
If y and z are jointly Gaussian random variables and $g(\cdot)$ is a nonlinear function, then $\mathbb{E}(g(y)z^*) = a\mathbb{E}(yz^*)$, where a is defined in (9).

This result allows us to characterize the linear MMSE fit for a Gaussian random vector in terms of the scalar linear MMSE fits for its components. It has been customized to MIMO in many recent papers [25], [27], [39], [40], hence we state the relevant result here without proof (see Appendix A in [40] for a derivation).

Theorem III.1. Vector Bussgang Decomposition
Let $\mathbf{y}$ *denote the jointly Gaussian random vector input to the effective nonlinearity* $g(\cdot)$ *referred to complex baseband, so that the received signal* $\mathbf{z} = g(\mathbf{y})$. *Then the Bussgang decomposition of* $\mathbf{z}$ *is given by*

$$\mathbf{z} = \mathbf{g}(\mathbf{y}) = \mathbf{A}\mathbf{y} + \mathbf{e}, \tag{12}$$

where

$$\mathbf{A} = \mathcal{D}iag([a_1, \ldots, a_N]), \tag{13}$$

$$a_i = \frac{\mathbb{E}(g(y_i)y_i^*)}{\mathbb{E}(|y_i|^2)}, \tag{14}$$

and the variance of element e_i *of the approximation error vector* $\mathbf{e}$ *is given by*

$$\sigma_{gi}^2 = \mathbb{E}(|g(y_i)|^2) - |a_i|^2\mathbb{E}(|y_i|^2). \tag{15}$$

The Bussgang theorem on covariance preservation therefore leads to a linear MMSE fit with diagonal covariance structure. Moreover, the diagonal elements are equal if the statistics of $\{y_i\}$ are identical, as in the following straightforward corollary, stated without proof.

Corollary 1. *If the diagonal elements of the covariance of* $\mathbf{y}$ *are equal, i.e.,* $\mathbb{E}(|y_i|^2) = \mathbb{E}(|y_k|^2)$, $\forall\, i, k$, *then the Bussgang decomposition specializes to*

$$\mathbf{z} = \mathbf{g}(\mathbf{y}) = a\mathbf{y} + \mathbf{e}, \tag{16}$$

where a *and* $\mathbb{E}(|e_i|^2) = \sigma_g^2$ *are the scalar Bussgang parameters of* $g(\cdot)$.

It is worth noting that the self-noise $\mathbf{e}$ may be spatially correlated. However, recent work [39] indicates that this correlation becomes negligible when the number of users is large, and we ignore it in our analysis here.

IV. Bussgang Normalization and Intrinsic SNR

In this section, we define a normalization such that the Bussgang parameters for a nonlinearity are independent of input power. We introduce the concept of *intrinsic SNR* to characterize the self-noise in this normalized setting. As we shall see, this is the summary specification that is provided by system-level design requirements to the hardware designer,

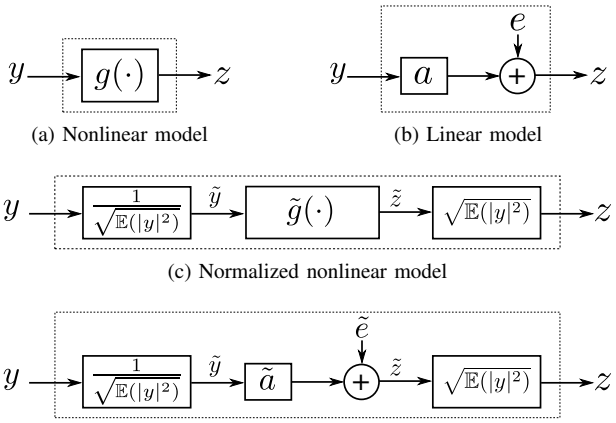


Figure 5: The nonlinear function $g(\cdot)$ in (a) can be decomposed to the linear model in (b) whose parameters depend on the input power. We define a normalized version of the nonlinearity in (c), which allows us to provide design specifications independent of input power. The corresponding normalized linearization is depicted in (d).

based on the analytical framework described in the next section. Finally, we show, via the simple example of a limiter, how such a summary can be used to determine hardware specifications for a nonlinearity.

Normalized Nonlinearity

As shown in Fig. 5 (a) and (b), Bussgang decomposition characterizes a nonlinear function $g(\cdot)$ by parameters a and σ_g^2. These parameters depend on the input power by definition as shown in Eq. (9) and (10).

Fig. 5 (c) illustrates a normalized version of the nonlinearity in Fig. 5 (a): the input power is scaled to one before the nonlinearity, and the scaling is undone after the nonlinearity. The Bussgang linearization of the normalized nonlinearity, with parameters $\tilde{a}$ and $\tilde{\sigma}_g^2$, is depicted in Fig. 5 (d). The parameters $\tilde{a}$ and $\tilde{\sigma}_g^2$ represent the Bussgang decomposition of the *normalized* nonlinear function $\tilde{g}(\cdot)$, depicted in Fig. 5 (c). The equivalence of the nonlinear models (a) and (c) implies that the corresponding linear models (b) and (d) must satisfy $\tilde{a} = a$ and $\tilde{\sigma}_g^2 = \sigma_g^2/\mathbb{E}(|y|^2)$.

It is convenient to define hardware specifications for the normalized nonlinearity; in hardware design parlance, the specifications are "referred to the input power." We summarize these using the concept of *intrinsic SNR,* which plays a key role in our analytical framework.

Definition IV.1. Intrinsic SNR
We define the "intrinsic SNR" of a nonlinearity $g(\cdot)$ using the Bussgang parameters of its normalized version $\tilde{g}(\cdot)$ as follows:

$$\gamma_g = \frac{|\tilde{a}|^2}{\tilde{\sigma}_g^2}. \tag{17}$$

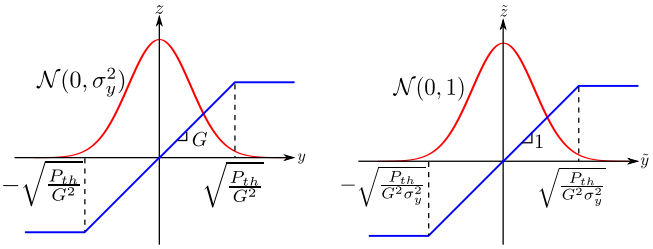


Figure 6: (a) The conventional limiter function. (b) a unity-gain limiter function whose clipping threshold is normalized to the effective input power.

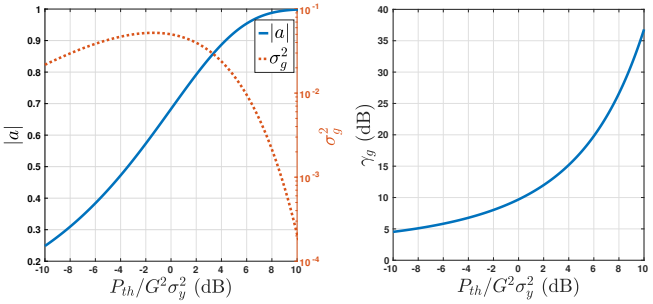


Figure 7: (a) Bussgang parameters and (b) the intrinsic SNR of the normalized limiter function.

As a simple example, consider a memoryless limiter as depicted in Fig. 6 (a), which is specified by the gain G and the power threshold P_{th} at which the output signal is clipped. The normalized version of this function has unity gain, as shown in Fig. 6 (b), hence we only need to specify a single parameter to characterize it: the clipping threshold $\tilde{P}_{th} = P_{th}/G^2\sigma_y^2$ normalized to the input power σ_y^2. The Bussgang parameters of the normalized limiter function are shown in Fig. 7 (a), and the intrinsic SNR is shown in Fig. 7 (b).

Henceforth, nonlinearities and their Bussgang parameters are normalized to the input power, and we drop the "tilde" notation to denote the normalized version. For example, the 1 dB compression point of a passband/baseband nonlinearity is normalized to the input power, and hence is measured in dB instead of dBm.

Design Approach

The analytical framework described in the next section leads the following design approach for going from system-level performance metrics to hardware design specifications:

• MIMO performance specifications lead to a requirement for the intrinsic SNR for the per-antenna nonlinearities, ignoring the specific nature of the nonlinearities. For example, suppose that we require an intrinsic SNR of 20 dB at least 95% of the

time.

• We map the intrinsic SNR requirement to a specification for the normalized nonlinearity. Taking the limiter in Fig. 6 as an example, we see from Fig. 7 (b), the clipping threshold normalized to the effective input power, $P_{th}/G^2\sigma_y^2$, must be at least 6 dB in order to attain an intrinsic SNR of 20 dB.

• In this step, the absolute value of the gain and clipping threshold is calculated. For example, suppose that the system in our running example is at load factor $\beta = 1/4$, i.e., 64 users. Then, according to the link budget presented in Appendix A, the input power to the receive chain is -60 dBm if power control is employed. We therefore obtain that $P_{th}/G^2 = -54$ dBm. The hardware designer now has to choose G and P_{th} in order to achieve this ratio or better.

V. Analytical Framework

Our analytical framework is developed as follows.

1) We derive a matched filter bound for each user in the MIMO system that accounts for the self-noise due to the per-antenna nonlinearities (which scales with the power summed across users) as well as thermal noise. To this end, we use Bussgang linearization and the intrinsic SNR discussed in the previous section.
2) We derive a lower bound for the output SINR of the LMMSE receiver for any given user. Defining the *efficiency* of the LMMSE receiver for a given user as the ratio of SINR to SNR, we show that the efficiency of a user in an ideal system without nonlinearities is a *lower bound* on that of the actual system. This, together with the matched filter bound, provides a lower bound on the LMMSE output SINR.
3) We obtain system-level design prescriptions by specializing the preceding lower bound to an "edge" user whose performance is stochastically poorer than that of any other user.

A. Bussgang Linearized Model

As described in section II, we denote by $\{A_k, k = 1, \ldots, K\}$ the amplitudes of the incoming waves for the K users, and by σ_n^2 the variance of the thermal noise at each antenna. We can therefore model the incoming signal at each receive antenna as $y_m \sim \mathcal{CN}(0, \sigma_y^2)$, where

$$\sigma_y^2 = \sum_{k=1}^{K} A_k^2 + \sigma_n^2 = \sigma_n^2 + KA_{\text{rms}}^2, \tag{18}$$

and

$$A_{\text{rms}} = \sqrt{\frac{1}{K}\sum_{k=1}^{K} A_k^2} \tag{19}$$

is the root mean square (rms) amplitude, averaged across users.

As depicted in Fig. 5 (c), using the normalized Bussgang linearization requires scaling the incoming signal to unit variance as follows:

$$\tilde{y}_m = \frac{y_m}{\sigma_y}. \tag{20}$$

For a normalized nonlinearity $g(\cdot)$ as defined in the previous section, our per antenna linearized model is given by:

$$g(\tilde{y}_m) = a\tilde{y}_m + e_m. \tag{21}$$

For the received signal (2), the normalized signal prior to passing through the nonlinearity is given by

$$\tilde{\mathbf{y}} = \frac{\mathbf{y}}{\sigma_y}. \tag{22}$$

Using the Bussgang decomposition, we have

$$g(\tilde{\mathbf{y}}) = a\tilde{\mathbf{y}} + \mathbf{e} = \frac{a}{\sigma_y}\mathbf{y} + \mathbf{e}, \tag{23}$$

where $\mathbb{E}[\mathbf{e}\mathbf{e}^H] = \sigma_g^2\mathbf{I}$. We can now go back to the original signal scaling to obtain

$$\hat{\mathbf{y}} = \frac{\sigma_y}{a}g(\tilde{\mathbf{y}}) = \mathbf{y} + \frac{\sigma_y}{a}\mathbf{e} = \mathbf{H}\mathbf{x} + \mathbf{n} + \frac{\sigma_y}{a}\mathbf{e}. \tag{24}$$

This is the model (5), with effective noise

$$\tilde{\mathbf{n}} = \mathbf{n} + \frac{\sigma_y}{a}\mathbf{e} \tag{25}$$

of variance

$$\mathbb{E}(\tilde{\mathbf{n}}\tilde{\mathbf{n}}^H) = (\sigma_n^2 + \nu_g^2)\mathbf{I} \tag{26}$$

where

$$\nu_g^2 = \frac{\sigma_y^2}{|a|^2}\sigma_g^2 = \frac{\sigma_y^2}{\gamma_g}. \tag{27}$$

B. Matched Filter Bound

For the k^{th} user, the matched filter bound for the linearized model (5), with equivalent noise as in (24)-(25), is simply given by

$$\text{SNR}_k(g) = \frac{||\mathbf{h}_k||^2}{\sigma_n^2 + \nu_g^2}. \tag{28}$$

Our design framework is built around the dependence of this bound on key system parameters as stated in the following theorem. We first ignore thermal noise, in order to clearly brings out the role of intrinsic SNR, γ_g, and load factor, β, and then include its effect.

Theorem V.1. Matched filter bound
(a) **Self-noise only:** *Ignoring thermal noise, the matched filter bound for user k is given by*

$$SNR_k(g) = \gamma_g \frac{A_k^2}{\beta A_{rms}^2}. \tag{29}$$

(b) **Self-noise and thermal noise:** *The matched filter bound for user k, considering both self-noise and thermal noise, is given by*

$$SNR_k(g, \sigma_n^2) = \frac{1}{\frac{1}{SNR_k(g)} + \frac{1+\gamma_g}{\gamma_g}\frac{1}{SNR_k}}, \tag{30}$$

where $SNR_k = NA_k^2/\sigma_n^2$ is the SNR for user k accounting for thermal noise alone.

Proof: The proof involves algebraic manipulations based on the linearized model (24)-(25).
(a) Using (1), the numerator in (28) is given by

$$||\mathbf{h}_k||^2 = NA_k^2. \tag{31}$$

Using (18) and (27), and setting $\sigma_n^2 = 0$, the denominator in (28) is given by

$$\nu_g^2 = \frac{KA_{\text{rms}}^2}{\gamma_g}. \tag{32}$$

Plugging (31) and (32) into (28), we obtain

$$\text{SNR}_k(g) = \frac{NA_k^2\gamma_g}{KA_{\text{rms}}^2} = \frac{\gamma_g A_k^2}{\beta A_{\text{rms}}^2}, \tag{33}$$

which is the desired result (29).
(b) From (28) and (31), we have

$$\frac{1}{\text{SNR}_k(g,\sigma_n^2)} = \frac{\sigma_n^2}{NA_k^2} + \frac{\nu_g^2}{NA_k^2}. \tag{34}$$

For non-zero thermal noise, we have, using (18) and (27), that

$$\nu_g^2 = \frac{KA_{\text{rms}}^2 + \sigma_n^2}{\gamma_g}. \tag{35}$$

Plugging into (34), we obtain upon simplification the desired result (30). ■

Note that, if $\gamma_g \gg 1$, then the formula (30) reduces to

$$\text{SNR}_k(g,\sigma_n^2) = \frac{1}{\frac{1}{\text{SNR}_k(g)} + \frac{1}{\text{SNR}_k}}. \tag{36}$$

In order to provide system-level performance guarantees, we focus on supporting users at the cell edge. We therefore now set A_k to the worst-case amplitude A_{edge} (at 100 m range for our running example), while computing A_{rms} by a statistical average $\sqrt{\mathbb{E}[A^2]}$ given the users distribution, assuming a large enough number of users. Let us term the ratio of the power of the edge user to the rms power as the *power control factor,* since it depends on the power control scheme used. The power control factor α_p is given by

$$\alpha_p = \frac{A_{\text{edge}}^2}{A_{\text{rms}}^2}. \tag{37}$$

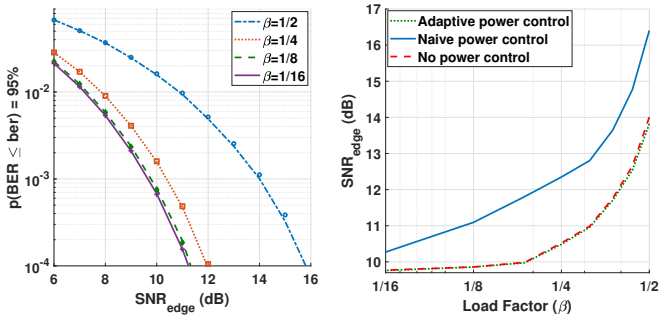
Specializing (29) to the edge user, we now obtain that

$$\text{SNR}_{\text{edge}}(g) = \gamma_g \frac{1}{\beta} \alpha_p. \tag{38}$$

Power control factor with no power control: For users who are uniformly distributed over the area bounded by $[R_{\text{min}}, R_{\text{max}}]$ and a given angular range, we obtain upon straightforward computation that, for a system without power control,

$$\begin{aligned}\alpha_p &= \frac{\frac{1}{R_{\text{max}}^2}}{\frac{1}{R_{\text{max}}^2 - R_{\text{min}}^2}\int_{R_{\text{min}}^2}^{R_{\text{max}}^2} \frac{1}{r} dr}, \\ &= \frac{1 - \frac{R_{\text{min}}^2}{R_{\text{max}}^2}}{2\log\frac{R_{\text{max}}}{R_{\text{min}}}}.\end{aligned} \tag{39}$$

which evaluates to -7.8 dB for $R_{\text{max}} = 100$ m, $R_{\text{min}} = 5$ m.



(a) BER in ideal system (b) SNR_{edge} required in ideal system

Figure 8: (a) BER for 5% outage in an ideal system (no nonlinearities) for different load factors. (b) SNR for an edge user (100 m from base station) to guarantee that 95% of the mobiles have raw BER of 10^{-3} for different load factors.

C. Lower Bound on LMMSE Output SINR

We now provide a lower bound on the output SINR of any user via the ideal system.

Theorem V.2. LMMSE Lower Bound
In the presence of nonlinearity, a lower bound on the output SINR of a linear MMSE receiver for any user is given as

$$SINR \geq SNR(g,\sigma_n^2)\ \eta_{ideal}, \tag{40}$$

where

$$\eta_{ideal} = \frac{SINR(ideal)}{SNR(ideal)}. \tag{41}$$

is the efficiency in an ideal system with the same user configuration and amplitudes, but without nonlinearity.

Proof: Since the system described in (5) is a pessimistic model for the system in (2), we have by Lemma C.1 in Appendix C that

$$\frac{\text{SINR}}{\text{SNR}(g,\sigma_n^2)} \geq \frac{\text{SINR(ideal)}}{\text{SNR(ideal)}}, \tag{42}$$

where SINR(ideal) is the target linear MMSE output SINR for a user in an ideal system (without nonlinearities). ■

We evaluate η_{ideal} through simulations of the ideal system for the edge user, as shown in Fig. 8, where $\text{SNR}_{\text{edge}} = \frac{NA_{\text{edge}}^2}{\sigma_n^2}$, and A_{edge} is the received amplitude of the user at 100 m. The target output SINR of the linear MMSE, i.e., $\text{SINR}_{\text{edge}}$(ideal) is 9.7 dB. This number corresponds to the SNR_{edge} in a single user case. Hence, in a single user case $\eta_{\text{ideal}} = 1$. As the load factor increases, there is noise enhancement due to interference suppression: $\eta_{\text{ideal}} = 9.7 - \text{SNR}_{\text{edge}}|_{\text{dB}}$ can be inferred from Fig. 8 (b).

D. From System-Level Performance to Intrinsic SNR

The chosen quality of service measure maps to an SINR requirement at the LMMSE output. We compute this for the

ideal system. For example, simulating the ideal system, a target BER of 10^{-3} with 95% availability is obtained for $\text{SINR}_{\text{edge}}(\text{ideal}) = 9.7$ dB. Since the SNR for an edge user is 14 dB, we see from Fig. 8 (b) that the efficiency for the ideal system is given by $9.7 - 14 = -4.3$ dB for no power control and $\beta = 1/2$. This is an upper bound on the efficiency of the actual system.

We can now compute the minimum $\text{SNR}_{\text{edge}}(g, \sigma_n^2)$ from Eq. (40) to achieve the required SINR in the presence of nonlinearities. Finally, we can infer the intrinsic SNR γ_g required from Eq. (30) and Eq. (36). This is now mapped to detailed hardware specifications, as illustrated by examples in the next section.

VI. Design Examples and Performance Evaluation

The system parameters are as described in Section II. We illustrate our design for a target uncoded BER of 10^{-3}, which is low enough for reliable performance using a high-rate channel code with relatively low decoding complexity. For QPSK, the corresponding required SNR over a SISO AWGN link is 9.7 dB. This becomes our target SINR at the output of the LMMSE receiver for an edge user. This setting is simply for illustration: our analytical framework applies for any QoS measure that can be approximated in terms of SINR (e.g., outage capacity or spectral efficiency using Shannon's formula).

In the following, we first describe the user distribution and power control schemes deployed in the cell. Then, we apply the analytical design framework to define the specification on the receive chain: the passband/baseband nonlinearity and the ADC resolution. We then evaluate the efficacy of the framework in attaining the desired system-level performance by simulations for selected scenarios. Finally, we provide design guidelines on the receive chain requirements in a more comprehensive set of scenarios.

A. User Distribution

The mobiles are uniformly distributed inside a region bordered by a minimum and a maximum distance away from the base station, R_{min} and R_{max}, respectively. Since $\frac{d\Omega}{d\theta} \sim \cos\theta$, the spatial frequency is less responsive to changes in DoA for θ near $\pm\frac{\pi}{2}$, which makes it more difficult to separate mobiles towards the edge of the angular field of view. We therefore confine the field of view for the antenna array to $-\pi/3 \leq \theta \leq \pi/3$. While the mobiles are placed randomly in our simulations, we enforce a minimum separation in spatial frequency between any two mobiles in order not to incur excessive interference, choosing it as half the 3 dB beamwidth: $\Delta\Omega_{\text{min}} = \frac{2.783}{N}$ [41] (mobiles closer in spatial frequency could be served in different time slots, for example). An example distribution of mobiles is depicted in Fig. 9.

B. Power Control Schemes

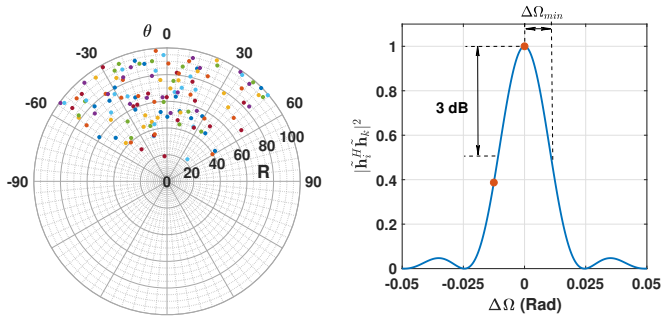
Our analysis in Section V first considers a system with no power control, in which each transmitter transmits at equal power. We then consider two power control schemes: a naive scheme in which transmitters adjust their powers to be roughly equal at the receiver, to within a tolerance, and an adaptive power control scheme designed for the linear MMSE receiver [42] aimed at meeting an SINR target for each mobile at the receiver. Power control is a very well-studied area, hence our goal is to provide quick insight on its implications for our system, rather than performing a comprehensive evaluation.

1) Naive power control: In this scheme, the base station asks all the users to decrease their power to make their received power at the base station equal the received power of the farthest mobile, i.e., at R_{max}. A disadvantage of this scheme, illustrated by our performance results in subsequent subsections, is that nearby users are no longer able to use their larger signal strength to overcome the impact of interference from other users who are nearby (in terms of spatial frequency separation). The power factor α_p of the naive power control scheme is equal to 0 dB because all the users have the same received signal strength.

2) Adaptive power control: In order to avoid the pitfalls of naive power control, we consider an adaptive power control scheme aimed at meeting an SINR target SINR_{th} at the output of the linear MMSE receiver. This method was proposed and shown to converge in [42]. We restate it with all values represented in the dB scale in Algorithm 1. Starting from no power control and all users transmitting at maximum power, the algorithm seeks to enforce a threshold SINR, termed SINR_{th}, iteratively as follows: every mobile with SINR greater than SINR_{th} reduces its power by $\text{SINR} - \text{SINR}_{\text{th}}$. This process is repeated until a convergence criterion is met. The power factor, α_p, that results from this adaptive power control scheme is found via simulation to equal about -2 dB for our running example.

C. Applying the Design Framework

For illustration, we consider $\beta = \frac{1}{2}$ and $\beta = \frac{1}{16}$, with no power control, naive power control and adaptive power control.



(a) Example distribution of mobiles (b) Normalized spatial cross-correlation

Figure 9: (a) An instantiation of 128 mobiles on a polar chart. (b) Normalized correlation between two users with spatial frequency difference of $\Delta\Omega$. Note that the closest users, depicted by red points, are separated by larger or equal to half the 3 dB beamwidth.

Algorithm 1: Adaptive power control

Input: $\mathbf{H}$, $\{P_k^{(0)}\}_{k=1,\ldots,K}$
parameter: SINR_{th}, ϵ
Output: $\{P_k\}_{k=1,\ldots,K}$

```
Margin = ε + 1;
while Margin > ε do
    {SINR_k}_{1:K} ← calculate LMMSE output SINR;
    for k ← 1 to K do
        ΔSINR_k ← max{SINR_k − SINR_th, 0};
        P_k ← P_k − ΔSINR_k;
    Margin ← max{ΔSINR_k}_{1:K};
```

The design steps are as follows:

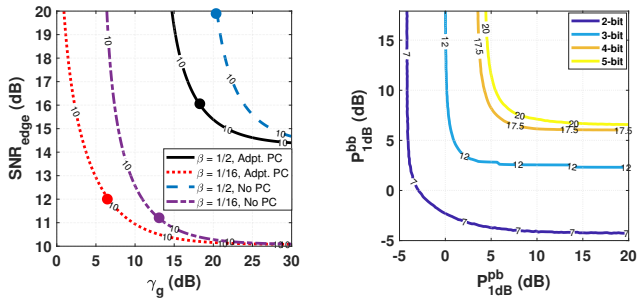
1) System-level design: We require $\text{SINR}_{\text{edge}}(\text{ideal}) \approx 10$ dB for our target QoS. Using simulations for the ideal system, we compute the LMMSE efficiency η_{ideal} as shown in Fig. 8. For our four scenarios, the LMMSE efficiency η_{ideal} is found to be (a) 4.5 dB, (b) 0 dB, (c) 4.5 dB, and (d) 0 dB. After that, we determine the SNR of the edge mobile and the intrinsic SNR jointly to achieve the LMMSE lower bound. Specifically, the contours in Fig. 10 (a) illustrates the following equation for each scenario:
$$\text{SNR}(g, \sigma_n^2) = \frac{\text{SINR}_{\text{edge}}}{\eta_{\text{ideal}}},$$
$$\frac{1}{\frac{\beta}{\gamma_g \alpha_p} + \frac{1+\gamma_g}{\gamma_g}\frac{1}{\text{SNR}_{\text{edge}}}} = \frac{10}{\eta_{\text{ideal}}}.$$
We pick the following combinations of (SNR_{edge},γ_g): (a) (20,20) dB, (b) (11,12) dB, (c) (16,17.5) dB, and (d) (12,7) dB.
2) Hardware-level design: This step determines the specifications of the passband/baseband nonlinearity and the ADC to achieve the required intrinsic SNR. Fig. 10 (b) shows the tradeoff between the number of ADC bits and the 1 dB compression point of the baseband nonlinearity $P_{\text{1dB}}^{\text{bb}}$ and the passband nonlinearity $P_{\text{1dB}}^{\text{pb}}$. The 1 dB compression points computed are normalized to the input power. The absolute compression points in dBm are computed by determining the average received input power at each base station antenna.

Here we have taken the link budget, or attainable SNR_{edge}, as our constraint, and have designed the nonlinearity specifications accordingly, We can, of course, utilize the same framework to determine the link budget required for a given set of nonlinearities.

D. Simulation-based Verification

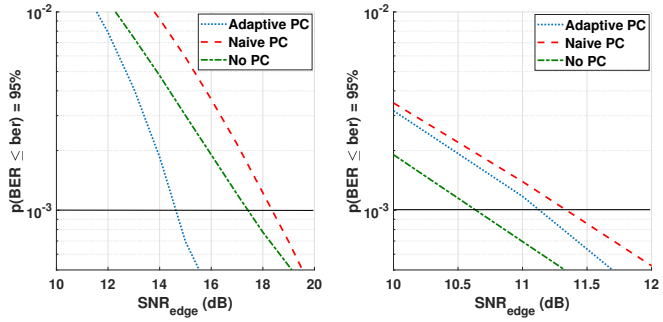
Here, we verify the designs produced by our analytical framework by numerical simulations. In Fig. 11, we plot the BER that 95% of the users attain for the cases we mention in the previous subsection. As shown, all the curves reach the 10^{-3} at slightly smaller SNR_{edge} than predicted by our analytical framework, which shows that our approach is both conservative and accurate. Table I summarizes our design prescriptions for different scenarios. As shown in the table, we examine the combination of four load factors with no power control and two power control strategies. We demonstrate the specification of the receive chain along with the resultant intrinsic SNR, γ_g. Then we compute an upper bound for the SNR needed for the edge user to achieve the performance metric. Finally, using simulations, we show the accuracy of the derived upper bound.

The results reported in Table I show that massive MIMO is key to relaxed hardware specifications: increasing the number of antennas for a given number of users reduces load factor, providing a "degrees of freedom" advantage that is used to level out the distortions caused by nonlinearities. For example,



(a) Contours of $\text{SINR}_{\text{edge}}$ (b) Contours of intrinsic SNR γ_g

Figure 10: (a) Lower bound on the linear MMSE output SINR as a function in the intrinsic SNR, γ_g, and the SNR required for the edge user, SNR_{edge}, for different scenarios. The contours depicted are for constant $\text{SINR}_{\text{edge}} = 10$ dB. The solid circles in Fig. (a) show the operating points we choose to work at. (b) Intrinsic SNR of a receive chain comprising passband and baseband nonlinearities and ADC.



(a) BER with load factor of 1/2 (b) BER with load factor of 1/16

Figure 11: (a) and (b) show the BER attained by 95% of the users for load factor of 1/2 and 1/16, respectively. The SNR_{edge} is the SNR required by the user at 100 m away from the base station. The receive chain specifications for each curve are demonstrated in table I.

Table I: This table presents the analytical predictions and simulation results for the SNR budget needed to meet the desired performance criterion (10^{-3} BER at 95% availability) for different scenarios. The intrinsic SNR γ_g corresponds to the cascade of the passband and baseband nonlinearities, specified by their 1 dB compression points ($P_{1\text{dB}}^{\text{pb}}$ and $P_{1\text{dB}}^{\text{bb}}$, respectively), together with b-bit ADCs for I and Q. PC and β denote the power control scheme used, and the load factor, respectively.

β	PC	b	$P_{1\text{dB}}^{\text{bb}}$ (dB)	$P_{1\text{dB}}^{\text{pb}}$ (dB)	γ_g (dB)	SNR_{edge} (upper bound)	SNR_{edge} (sim.)
1/2	none	5	8.4	6.7	20	20	17.5
1/2	naive	4	8.4	4.9	17.5	18.7	18.4
1/2	adaptive	4	8.4	4.9	17.5	16	14.7
1/4	none	4	8.2	2.4	15	14	12.8
1/4	naive	3	3.7	0.7	10.5	15	14.8
1/4	adaptive	3	3.4	1.9	11.5	12.5	11.9
1/8	none	3	4.2	1.4	12	13	12.2
1/8	naive	3	2.2	-1.1	8.7	12.7	12.7
1/8	adaptive	3	3.2	1.9	11	10.9	10.8
1/16	none	3	4.2	1.4	12	11.2	10.8
1/16	naive	2	1.4	-1.1	7.6	11.5	11.5
1/16	adaptive	2	-1.1	-1.9	7	11.8	11.2

when serving 16 users with a 256 element array ($\beta = \frac{1}{16}$), only 2 bits of quantization per dimension is required and the 1-dB compression point of RF and baseband amplifiers is very low (both below 0 dB with adaptive power control). In practice, a lower compression point allows higher loading, i.e., the amplifier can support a larger input signal and produce a stronger signal in the output, which increases the power efficiency of an amplifier. This is very desirable trait for scaling to large arrays. The results also bring out the impact of power control. Higher disparities in user powers requires higher ADC granularity to allow effective interference suppression of strong users which might otherwise "drown out" weak users. The reduction in user power disparities and required dynamic range with better power control is reflected clearly in our results.

VII. Conclusion

The analytical framework provided in this paper is a conservative, yet accurate, approach for designing hardware specifications for nonlinear elements in all-digital mmWave massive MIMO. Scaling using a larger number of antennas with a smaller load factor is attractive, since the specifications for RF nonlinearities, baseband nonlinearities, and ADC precision can all be significantly relaxed by operating at lower load factors. The requirements can also be relaxed by use of appropriate power control, as illustrated by the simple adaptive power control scheme considered here.

While we have considered LoS channel models here, we note that our approach extends to sparse multipath channels. At high symbol rates, equalization over a large delay spread becomes computationally unattractive. In this case paths that differ significantly in delay and angular spread from the dominant path play the role of additional interference, and can be folded into our framework.

In addition to the extensive effort required to realize our design prescriptions in hardware, there are also important open issues related to the digital backend, given the challenges of both computation and data transport for the multiGigabaud, multiuser system considered here. Thus, despite the extensive prior research on multiuser detection, there are significant open issues on the design of strategies that are efficient enough (in terms of both computation and communication on the backend fabric) to scale with the number of antennas, number of users, and bandwidth. We also note the importance of exploring *nonlinear* reception techniques (such as interference cancellation) that could outperform LMMSE detection while maintaining low computational overhead. Preliminary results in [43]–[45] indicate that exploiting channel sparsity is a promising approach for developing such techniques.

Appendix A
All-digital link budget

We provide here example parameters that demonstrate that the link budget for all-digital massive multiuser MIMO uplink system is realizable with low-cost silicon:

- antenna element gain covering a hemisphere is 3 dBi,
- 16-element array at the mobile gives 12 dBi transmit beamforming gain, plus 12 dB power pooling gain,
- 256-element array in the base station gives 24 dBi receive beamforming gain,
- noise figure for each RF chain in the base station of 7 dB,
- thermal noise power over 5 GHz bandwidth is about -77 dBm,
- and free space path loss of an edge user at 100 m using a carrier frequency of 140 GHz is about 115 dB.

The transmit power required from each power amplifier (PA) at the mobile to achieve a target SNR (in dB) for an edge-user, namely $\text{SNR}_{\text{edge}}|_{\text{dB}}$, can now be computed as

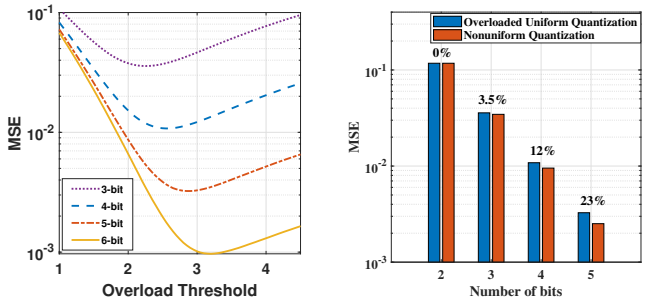
$$P_{\text{PA}} = \text{SNR}_{\text{edge}}|_{\text{dB}} - 9 \ \ \text{dBm}. \tag{43}$$

For example, $\text{SNR}_{\text{edge}}|_{\text{dB}}$ of about 16 dB (shown to suffice for our case study) requires 7 dBm PA output, which is realizable in CMOS (CMOS designs of up to 11 dBm have been reported in [46]).

Appendix B
Uniform vs nonuniform quantization

Our simulation results are for an overloaded ADC. The overloaded uniform ADC comprises two regions in its I/O characteristic, the granular and overload regions. The granular region is quantized uniformly, with bounded quantization noise. While quantization noise in the overload region, represented by the quantizer levels at the edges, is unbounded, the contribution to the MSE is kept comparable to that of the granular region by minimizing the MSE for the given input distribution; see Fig. 12 (a), where MSE is plotted against overload threshold.

An alternative is to employ an MSE-optimal quantizer using Lloyd's algorithm [47], with quantization bins as listed in [48]. The MSE comparison between these two options is shown in Fig. 12 (b). The advantage of nonuniform MSE-optimal



(a) (b)

Figure 12: (a) MSE versus overload threshold. (b) MSE comparison of overload uniform quantizer versus MSE-optimal nonuniform quantizer. The percentages represent the relative reduction in MSE from using MSE-optimal nonuniform quantization

quantization is barely noticeable for the small number of quantization bits of interest here, hence we choose to work with the simpler overloaded uniform quantizer.

Appendix C
Linear MMSE properties

From the point of view of a given user (the *desired* user) with channel $\mathbf{h}$, we may write the received signal corresponding to a single symbol as

$$\mathbf{r} = s\mathbf{h} + \mathbf{w}_I + \mathbf{w}_N, \tag{44}$$

where s denotes the transmitted symbol, $\mathbf{w}_I$ denotes the interference vector and $\mathbf{w}_N$ is the zero mean noise vector with covariance matrix $\sigma_n^2\mathbf{I}$. Standard assumptions necessary for effective interference suppression are that the desired symbol is uncorrelated with the interference and noise: $\mathbb{E}[s^*\mathbf{w}_I] = \mathbb{E}[s^*\mathbf{w}_N] = \mathbf{0}$. We also assume that the interference and noise are uncorrelated.

A linear correlator $\mathbf{c}$ produces a decision statistic $\mathbf{c}^H\mathbf{r}$ for the desired symbol, and its SINR is given by

$$\begin{aligned}\text{SINR}(\mathbf{c}) =& \frac{\mathbb{E}[|s\mathbf{c}^H\mathbf{h}|^2]}{\mathbb{E}[|\mathbf{c}^H(\mathbf{w}_I+\mathbf{w}_N)|^2]} \\ =& \frac{\sigma_s^2|\mathbf{c}^H\mathbf{h}|^2}{\mathbf{c}^H\mathbf{R}_I\mathbf{c}+\sigma_n^2||\mathbf{c}||^2},\end{aligned} \tag{45}$$

where $\mathbf{R}_I = \mathbb{E}[\mathbf{w}_I\mathbf{w}_I^H]$ is the interference covariance matrix, and $\mathbf{R}_N = \mathbb{E}[\mathbf{w}_N\mathbf{w}_N^H] = \sigma_n^2\mathbf{I}$ is the noise covariance matrix.

The LMMSE correlator minimizes MSE $= \mathbb{E}[|\mathbf{c}^H\mathbf{r} - s|^2]$ and maximizes SINR [35]. For the additive noise-plus-interference model (44), it is known to be proportional to a whitened matched filter (i.e., it suppresses interference by whitening it):

$$\mathbf{c}_{\text{MMSE}} = \alpha(\mathbf{R}_I+\mathbf{R}_N)^{-1}\mathbf{h} = \alpha(\mathbf{R}_I+\sigma_n^2\mathbf{I})^{-1}\mathbf{h}, \tag{46}$$

where α is a scale factor that can be solved for easily (e.g., see [35]). Since SINR does not depend on scale factor, it is easy to show, plugging into (45), that

$$\begin{aligned}\text{SINR} =& \sigma_s^2\mathbf{h}^H\left(\mathbf{R}_I+\mathbf{R}_N\right)^{-1}\mathbf{h} \\ =& \sigma_s^2\mathbf{h}^H\left(\mathbf{R}_I+\sigma_n^2\mathbf{I}\right)^{-1}\mathbf{h}.\end{aligned} \tag{47}$$

Let us also for reference define the SNR:

$$\text{SNR} = \sigma_s^2\mathbf{h}^H\left(\mathbf{R}_N\right)^{-1}\mathbf{h} = \sigma_s^2||\mathbf{h}||^2/\sigma_n^2. \tag{48}$$

Remark 1. *A positive definite matrix $\mathbf{A}(\theta)$ increases with θ if $\mathbf{A}(\theta) - \mathbf{A}(\theta') \geq \mathbf{0}$ for any $\theta > \theta'$. That is, for any vector $\mathbf{u}$, $\mathbf{u}^H\mathbf{A}(\theta)\mathbf{u} \geq \mathbf{u}^H\mathbf{A}(\theta')\mathbf{u}$.*

We can now infer the following properties relevant for our approach to performance analysis, stated as a lemma.

Lemma C.1. *If the noise level σ_n^2 increases, with the signal and interference characteristics unchanged, then*
(a) Absolute performance gets worse, with SINR and SNR both decreasing.
(b) The noise enhancement gets better: $\frac{SNR}{SINR}$ decreases.

Proof: For (a), we note that the positive definite matrix $\mathbf{R}_I + \sigma_n^2\mathbf{I}$ increases with σ_n^2, hence its inverse decreases with σ_n^2. For (b), note that

$$\begin{aligned}\frac{\text{SNR}}{\text{SINR}} =& \frac{||\mathbf{h}||^2/\sigma_n^2}{\mathbf{h}^H\left(\mathbf{R}_I+\sigma_n^2\mathbf{I}\right)^{-1}\mathbf{h}}, \\ =& \frac{||\mathbf{h}||^2}{\mathbf{h}^H\left(\mathbf{R}_I/\sigma_n^2+\mathbf{I}\right)^{-1}\mathbf{h}}.\end{aligned} \tag{49}$$

The positive definite matrix $\mathbf{R}_I/\sigma_n^2 + \mathbf{I}$ decreases with σ_n^2, hence its inverse increases with σ_n^2. Thus, the denominator on the right-hand side of equation (49) increases with σ_n^2, while the numerator is independent of it, proving the desired result. ■

Acknowledgments

This work was supported in part by the Semiconductor Research Corporation (SRC) under the JUMP program (2018-JU-2778) and by DARPA (HR0011-18-3-0004). Use was made of the computational facilities administered by the Center for Scientific Computing at the CNSI and MRL (an NSF MRSEC; DMR-1720256) and purchased through NSF CNS-1725797.

References

[1] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive mimo for next generation wireless systems," *IEEE communications magazine*, vol. 52, no. 2, pp. 186–195, 2014.

[2] V. Jungnickel, K. Manolakis, W. Zirwas, B. Panzner, V. Braun, M. Lossow, M. Sternad, R. Apelfröjd, and T. Svensson, "The role of small cells, coordinated multipoint, and massive mimo in 5g," *IEEE communications magazine*, vol. 52, no. 5, pp. 44–51, 2014.

[3] A. Puglielli, A. Townley, G. LaCaille, V. Milovanović, P. Lu, K. Trotskovsky, A. Whitcombe, N. Narevsky, G. Wright, T. Courtade *et al.*, "Design of energy-and cost-efficient massive mimo arrays," *Proceedings of the IEEE*, vol. 104, no. 3, pp. 586–606, 2015.

[4] A. Babakhani, X. Guan, A. Komijani, A. Natarajan, and A. Hajimiri, "A 77-ghz phased-array transceiver with on-chip antennas in silicon: Receiver and antennas," *IEEE Journal of Solid-State Circuits*, vol. 41, no. 12, pp. 2795–2806, 2006.

[5] W. Hong, K.-H. Baek, Y. Lee, Y. Kim, and S.-T. Ko, "Study and prototyping of practically large-scale mmwave antenna systems for 5g cellular devices," *IEEE Communications Magazine*, vol. 52, no. 9, pp. 63–69, 2014.

[6] E. Cohen, M. Ruberto, M. Cohen, O. Degani, S. Ravid, and D. Ritter, "A cmos bidirectional 32-element phased-array transceiver at 60 ghz with ltcc antenna," *IEEE Transactions on Microwave Theory and Techniques*, vol. 61, no. 3, pp. 1359–1375, 2013.

[7] A. Valdes-Garcia, S. T. Nicolson, J.-W. Lai, A. Natarajan, P.-Y. Chen, S. K. Reynolds, J.-H. C. Zhan, D. G. Kam, D. Liu, and B. Floyd, "A fully integrated 16-element phased-array transmitter in sige bicmos for 60-ghz communications," *IEEE journal of solid-state circuits*, vol. 45, no. 12, pp. 2757–2773, 2010.

[8] Z. Marzi, D. Ramasamy, and U. Madhow, "Compressive channel estimation and tracking for large arrays in mm-wave picocells," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 514–527, 2016.

[9] H. Yan and D. Cabric, "Compressive initial access and beamforming training for millimeter-wave cellular systems," *IEEE journal of selected topics in signal processing*, vol. 13, no. 5, pp. 1151–1166, 2019.

[10] F. Sohrabi and W. Yu, "Hybrid digital and analog beamforming design for large-scale antenna arrays," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 501–513, 2016.

[11] O. Bakr, M. Johnson, J. Park, E. Adabi, K. Jones, and A. Niknejad, "A scalable-low cost architecture for high gain beamforming antennas," in *2010 IEEE International Symposium on Phased Array Systems and Technology*. IEEE, 2010, pp. 806–813.

[12] G. Zhu, K. Huang, V. K. Lau, B. Xia, X. Li, and S. Zhang, "Hybrid beamforming via the kronecker decomposition for the millimeter-wave massive mimo systems," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 2097–2114, 2017.

[13] J.-C. Chen, "Hybrid beamforming with discrete phase shifters for millimeter-wave massive mimo systems," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 8, pp. 7604–7608, 2017.

[14] M. Chung, L. Liu, O. Edfors, and F. Tufvesson, "Millimeter-wave massive mimo testbed with hybrid beamforming," in *2020 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*. IEEE, 2020, pp. 1–2.

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

[16] H. Yan, S. Ramesh, T. Gallagher, C. Ling, and D. Cabric, "Performance, power, and area design trade-offs in millimeter-wave transmitter beamforming architectures," *IEEE Circuits and Systems Magazine*, vol. 19, no. 2, 2019.

[17] O. T. Demir and E. Bjornson, "The bussgang decomposition of nonlinear systems: Basic theory and mimo extensions [lecture notes]," *IEEE Signal Processing Magazine*, vol. 38, no. 1, pp. 131–136, 2020.

[18] B. Razavi and R. Behzad, *RF microelectronics*. Prentice Hall New Jersey, 1998, vol. 2.

[19] J. BUSSGANG, "Crosscorrelation functions of amplitude-distorted gaussian signals," *MIT Res. Lab. Elec. Tech. Rep.*, vol. 216, 1952.

[20] E. Björnson, J. Hoydis, M. Kountouris, and M. Debbah, "Massive MIMO systems with non-ideal hardware: Energy efficiency, estimation, and capacity limits," *IEEE Transactions on Information Theory*, vol. 60, no. 11, 2014.

[21] L. Fan, S. Jin, C.-K. Wen, and H. Zhang, "Uplink achievable rate for massive MIMO systems with low-resolution ADC," *IEEE Communications Letters*, vol. 19, no. 12, 2015.

[22] J. Zhang, L. Dai, S. Sun, and Z. Wang, "On the spectral efficiency of massive MIMO systems with low-resolution ADCs," *IEEE Communications Letters*, vol. 20, no. 5, 2016.

[23] L. Xu, X. Lu, S. Jin, F. Gao, and Y. Zhu, "On the uplink achievable rate of massive MIMO system with low-resolution ADC and RF impairments," *IEEE Communications Letters*, vol. 23, no. 3, 2019.

[24] C. Mollén, J. Choi, E. G. Larsson, and R. W. Heath, "Achievable uplink rates for massive MIMO with coarse quantization," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017.

[25] S. Jacobsson, G. Durisi, M. Coldrey, U. Gustavsson, and C. Studer, "Throughput analysis of massive MIMO uplink with low-resolution ADCs," *IEEE Transactions on Wireless Communications*, vol. 16, no. 6, 2017.

[26] C. Studer and G. Durisi, "Quantized massive MU-MIMO-OFDM uplink," *IEEE Transactions on Communications*, vol. 64, no. 6, 2016.

[27] S. Jacobsson, U. Gustavsson, G. Durisi, and C. Studer, "Massive MU-MIMO-OFDM uplink with hardware impairments: Modeling and analysis," in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2018.

[28] A. Maltsev, A. Pudeyev, I. Karls, I. Bolotin, G. Morozov, R. Weiler, M. Peter, and W. Keusgen, "Quasi-deterministic approach to mmWave channel modeling in a non-stationary environment," in *2014 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2014.

[29] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE access*, vol. 1, 2013.

[30] T. S. Rappaport, F. Gutierrez, E. Ben-Dor, J. N. Murdock, Y. Qiao, and J. I. Tamir, "Broadband millimeter-wave propagation measurements and models using adaptive-beam antennas for outdoor urban cellular communications," *IEEE transactions on antennas and propagation*, vol. 61, no. 4, 2012.

[31] M. Jacob, S. Priebe, R. Dickhoff, T. Kleine-Ostmann, T. Schrader, and T. Kurner, "Diffraction in mm and sub-mm wave indoor propagation channels," *IEEE Transactions on Microwave Theory and Techniques*, vol. 60, no. 3, 2012.

[32] M. Abdelghany, A. Farid, U. Madhow, and M. Rodwell, "Towards all-digital mmWave massive MIMO: Designing around nonlinearities," in *Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2018.

[33] D. C. Montgomery and G. C. Runger, *Applied statistics and probability for engineers*. Wiley, 2014.

[34] A. Gersho and R. M. Gray, *Vector quantization and signal compression*. Springer Science & Business Media, 2012, vol. 159.

[35] U. Madhow and M. L. Honig, "MMSE interference suppression for direct-sequence spread-spectrum CDMA," *IEEE transactions on communications*, vol. 42, no. 12, 1994.

[36] S. Verdu *et al.*, *Multiuser detection*. Cambridge university press, 1998.

[37] B. Hajek, *Random processes for engineers*. Cambridge university press, 2015.

[38] J. Minkoff, "The role of AM-to-PM conversion in memoryless nonlinear systems," *IEEE Transactions on Communications*, vol. 33, no. 2, 1985.

[39] E. Björnson, L. Sanguinetti, and J. Hoydis, "Hardware distortion correlation has negligible impact on UL massive MIMO spectral efficiency," *IEEE Transactions on Communications*, vol. 67, no. 2, 2018.

[40] S. Jacobsson, G. Durisi, M. Coldrey, T. Goldstein, and C. Studer, "Quantized precoding for massive MU-MIMO," *IEEE Transactions on Communications*, vol. 65, no. 11, 2017.

[41] C. A. Balanis, *Antenna Theory: Analysis and Design*. New York, NY, USA: Wiley-Interscience, 2005.

[42] S. Ulukus and R. D. Yates, "Adaptive power control and MMSE interference suppression," *Wireless Networks*, vol. 4, no. 6, 1998.

[43] M. Abdelghany, U. Madhow, and A. Tölli, "Beamspace local LMMSE: An efficient digital backend for mmWave massive MIMO," in *2019*

IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC). IEEE, 2019, pp. 1–5.

[44] M. Abdelghany, U. Madhow, and M. Rodwell, "An efficient digital backend for wideband single-carrier mmWave massive MIMO," in *to be presented in IEEE Global Communications Conference (Globecom), Waikoloa, Hawaii, Dec. 2019*, 2019.

[45] M. Abdelghany, M. E. Rasekh, and U. Madhow, "Scalable nonlinear multiuser detection for mmwave massive mimo," in *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2020, pp. 1–5.

[46] D. Simic and P. Reynaert, "A 14.8 dBm 20.3 dB power amplifier for D-band applications in 40 nm CMOS," in *2018 IEEE Radio Frequency Integrated Circuits Symposium (RFIC)*. IEEE, 2018.

[47] S. Lloyd, "Least squares quantization in PCM," *IEEE transactions on information theory*, vol. 28, no. 2, 1982.

[48] J. Max, "Quantizing for minimum distortion," *IRE Transactions on Information Theory*, vol. 6, no. 1, 1960.

Mohammed Abdelghany received his B.Sc. and M.Sc. degrees in Electrical Communications Engineering from Cairo University, Cairo, Egypt, in 2012 and 2016, respectively. In 2015, he was a visiting student with Prof. Katabi at MIT Computer Science & Artificial Intelligence Lab, Cambridge, USA. He is currently pursuing his Ph.D. in the Wireless Communication and Sensornets Lab at UCSB, Santa Barbara, USA. His research interests include signal processing, wireless communications, and design of digital VLSI circuits and systems.

Ali Farid received his B.Sc and M.Sc degrees in Electronics and Communications Engineering from Cairo University, Cairo, Egypt in 2012 and 2016, respectively. From 2012 to 2016 he was a full time Analog/RF IC design engineer at Atmel corporation, Egypt, where he was involved in the design of high resolution data converters (ADCs/DACs) and low power transceiver frontends for connectivity solutions. He is currently pursuing his PhD at University of California, Santa Barbara, CA, USA. His current research interests include mmWave/RF systems, circuits and modules design for multiuser massive MIMO arrays and single beam phased arrays.

Maryam Eslami Rasekh received her B.S. and M.S. degrees in Electrical Engineering from Isfahan University of Technology in 2007 and Sharif University of Technology in 2009, respectively. She joined the Wireless Communication and Sensornets Lab at University of California Santa Barbara as a visiting scholar in 2013 and received her Ph.D. from UC Santa Barbara in 2020. She is currently a postdoctoral researcher at UC Santa Barbara working with Prof. Madhow in the WCSL lab. Her research is mainly focused on millimeter wave communication, networking, and sensing, with emphasis on next generation networks and applications. Her past and current research in this area includes wave propagation and channel modeling, compressive channel estimation and beamforming, backhaul network optimization, signal processing and frontend design for multiuser massive MIMO systems, and massive MIMO radar.

Upamanyu Madhow is Distinguished Professor of Electrical and Computer Engineering at the University of California, Santa Barbara. His current research interests focus on next generation communication, sensing and inference infrastructures centered around millimeter wave systems, and on fundamentals and applications of robust machine learning. He received his bachelor's degree in electrical engineering from the Indian Institute of Technology, Kanpur, in 1985, and his Ph.D. degree in electrical engineering from the University of Illinois, Urbana-Champaign in 1990. He has worked as a research scientist at Bell Communications Research, Morristown, NJ, and as a faculty at the University of Illinois, Urbana-Champaign. Dr. Madhow is a recipient of the 1996 NSF CAREER award, and co-recipient of the 2012 IEEE Marconi prize paper award in wireless communications. He has served as Associate Editor for the IEEE Transactions on Communications, the IEEE Transactions on Information Theory, and the IEEE Transactions on Information Forensics and Security. He is the author of two textbooks published by Cambridge University Press, Fundamentals of Digital Communication (2008) and Introduction to Communication Systems (2014).

Mark Rodwell holds the Doluca Family Endowed Chair in Electrical and Computer Engineering at UCSB and directs the SRC/DARPA Center for Converged TeraHertz Communications and Sensing. His group develops nm and THz transistors, high-frequency ICs, and wireless communcations systems. He and his collaborators received the 2010 IEEE Sarnoff Award, the 2012 Marconi Prize Paper Award, the 1997 IEEE Microwave Prize, the 2009 IEEE IPRM Conference Award, and the 1998 European Microwave Conference Microwave Prize.