1-bit Massive MU-MIMO Precoding in VLSI

Oscar Castañeda, Sven Jacobsson, Giuseppe Durisi,
Mikael Coldrey, Tom Goldstein, and Christoph Studer

***Abstract*—Massive multiuser (MU) multiple-input multiple-output (MIMO) will be a core technology in fifth-generation (5G) wireless systems as it offers significant improvements in spectral efficiency compared to existing multi-antenna technologies. The presence of hundreds of antenna elements at the base station (BS), however, results in excessively high hardware costs and power consumption, and requires high interconnect throughput between the baseband-processing unit and the radio unit. Massive MU-MIMO that uses low-resolution analog-to-digital and digital-to-analog converters (DACs) has the potential to address all these issues. In this paper, we focus on downlink precoding for massive MU-MIMO systems with 1-bit DACs at the BS. The objective is to design precoders that simultaneously mitigate multi-user interference (MUI) and quantization artifacts. We propose two nonlinear 1-bit precoding algorithms and corresponding very-large scale integration (VLSI) designs. Our algorithms rely on biconvex relaxation, which enables the design of efficient 1-bit precoding algorithms that achieve superior error-rate performance compared to that of linear precoding algorithms followed by quantization. To showcase the efficacy of our algorithms, we design VLSI architectures that enable efficient 1-bit precoding for massive MU-MIMO systems in which hundreds of antennas serve tens of user equipments. We present corresponding field-programmable gate array (FPGA) implementations to demonstrate that 1-bit precoding enables reliable and high-rate downlink data transmission in practical systems.**

***Index Terms*—Biconvex relaxation, digital-to-analog converter (DAC), field-programmable gate array (FPGA), massive multi-user multiple-input multiple-output (MU-MIMO), precoding, quantization, very large-scale integration (VLSI).**

I. Introduction

MASSIVE multiuser (MU) multiple-input multiple-output (MIMO) is widely believed to be a core technology in fifth-generation (5G) wireless systems as it enables substantial improvements in spectral efficiency and reliability compared to traditional, small-scale MIMO technology [2]–[4]. These advantages are a result of equipping the base station (BS) with hundreds or thousands of antennas, which enables fine-grained beamforming to serve tens of user equipments (UEs) in the same time-frequency resource. However, the large number of antenna elements and radio frequency (RF) chains at the BS results in a significant increase in hardware complexity, system costs, and circuit power consumption. Furthermore, massive MU-MIMO requires high interconnect and chip input/output (I/O) bandwidth between the baseband-processing unit at the BS and the radio units [5]. As a consequence, the successful deployment of this technology in 5G wireless systems requires novel design approaches that jointly reduce system costs, power consumption, and interconnect bandwidth without degrading the spectral efficiency and reliability.

A. Massive MU-MIMO with 1-bit DACs

We consider the massive MU-MIMO downlink in which the BS is equipped with 1-bit digital-to-analog converters (DACs) and transmits data to multiple UEs in the same time-frequency resource. In traditional multi-antenna BSs, each RF port is connected to a pair of high-resolution DACs (e.g., with 10-bit precision). Scaling such architectures to massive MIMO BSs, with hundreds or thousands of antennas would result in prohibitively high power consumption and system costs. The deployment of 1-bit DACs at the BS mitigates this problem. In addition, the use of 1-bit DACs enables one to lower the linearity and noise requirements of the surrounding RF circuitry, which has the potential to additionally reduce the RF circuit power. Another benefit of using 1-bit DACs is the fact that lowering their resolution also reduces the fronthaul bandwidth between the baseband-processing unit and the radio unit. This aspect is of practical relevance for deployment scenarios in which these two units are separated by a large distance.

The key challenges of 1-bit massive MU-MIMO systems are to maintain high spectral efficiency and reliability. As shown in [1], [6]–[9], the use of 1-bit DACs in the downlink enables reliable data transmission if sophisticated precoding algorithms that simultaneously mitigate multi-user interference (MUI) and quantization artifacts are used. While conventional linear precoding methods, such as zero-forcing (ZF) or minimum mean-square error (MMSE) precoding followed by quantization, require low computational complexity [10]–[13], more sophisticated, nonlinear methods are necessary to enable reliable communication at high spectral efficiency. Such precoding methods, however, typically require high computational complexity. As a consequence, a successful deployment of 1-bit massive MU-MIMO calls for the design of novel and efficient precoding algorithms that can be implemented in hardware and reliably achieve high throughput.

O. Castañeda and C. Studer are with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY (e-mail: oc66@cornell.edu, studer@cornell.edu).

S. Jacobsson is with Ericsson Research and Chalmers University of Technology, Gothenburg, Sweden (e-mail: sven.jacobsson@ericsson.com).

G. Durisi is with Chalmers University of Technology, Gothenburg, Sweden (e-mail: durisi@chalmers.se).

M. Coldrey is with Ericsson Research, Gothenburg, Sweden (e-mail: mikael.coldrey@ericsson.com)

T. Goldstein is with the Department of Computer Science, University of Maryland, College Park, MD (e-mail: tomg@cs.umd.edu).

The C1PO algorithm implemented in this paper builds upon the 1-bit precoding algorithm to be presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing [1]; in contrast to the algorithm in [1], C1PO directly operates in the complex domain, comes with convergence guarantees, and can be implemented efficiently in VLSI.

A system simulator for 1-bit precoding in massive MU-MIMO systems will be made available on GitHub after (possible) acceptance of the paper.

B. Contributions

In this paper, we develop novel, computationally efficient precoding algorithms for 1-bit massive MU-MIMO systems and corresponding very-large scale integrated (VLSI) designs. Our main contributions can be summarized as follows:

- We use biconvex relaxation (BCR) [14] to design a nonlinear 1-bit precoding algorithm. Our algorithm, referred to as C1PO (short for biConvex 1-bit PrecOding), enables reliable, high-rate downlink transmission in 1-bit massive MU-MIMO systems for medium-sized antenna arrays.
- We propose a scalable and low-complexity algorithm variant, referred to as C2PO, which enables high-performance 1-bit precoding for massive MU-MIMO systems with hundreds or thousands of antenna elements.
- For C1PO and C2PO, which both solve nonconvex problems, we provide analytical convergence guarantees.
- We develop two massively parallel VLSI architectures that implement C1PO and C2PO, and achieve high throughputs in a hardware-efficient manner. Our architectures support various BS and UE antenna configurations.
- We present reference designs on a Xilinx Virtex-7 field-programmable gate array (FPGA) for various antenna configurations that demonstrate the efficacy of our algorithms and VLSI architectures.
- We compare our designs to a baseline precoder that uses maximum ratio combining (MRC) followed by quantization (MRC-Q), a method that achieves high hardware-efficiency at the cost of poor error-rate performance.
- We study the trade-offs between error-rate performance and hardware efficiency (in terms of throughput per area) for the proposed FPGA designs.

Our results demonstrate the practical feasibility of 1-bit precoding in massive MU-MIMO systems, supporting reliable and high-rate downlink data transmission.

C. Relevant Prior Art

A number of papers have studied the use of low-resolution analog-to-digital converters (ADCs) for the massive MU-MIMO uplink (UEs transmit data to the BS) with particular focus on the 1-bit case; see, e.g., [15]–[19] and the references therein. All these results have shown that the use of 1-bit ADCs is sufficient for reliable low-rate uplink transmission and that 4-to-6 bits are sufficient to close the gap to the infinite-precision case in most scenarios. In contrast to the uplink, the quantized downlink has gained attention only recently. The results in [10]–[13] have shown that so-called *linear-quantized* precoders, which perform traditional linear precoding followed by quantization, enable reliable uplink transmission for very large BS antenna arrays in the high signal-to-noise ratio (SNR) regime, even for systems that use 1-bit DACs. More sophisticated, *nonlinear* precoding algorithms have been proposed only recently in [1], [6]–[9] and significantly outperform linear-quantized methods in the case of 1-bit DACs. The computational complexity of these algorithms, however, is typically high, which prevents an efficient implementation in practical systems. In contrast to these precoding methods, we propose two novel nonlinear precoding algorithms and VLSI designs that achieve high throughput in a hardware-efficient manner.

While a large number of VLSI designs for data detection in the massive MU-MIMO uplink have been proposed in the literature (see, e.g., [20]–[23] and references therein), only a handful of precoder designs for multi-antenna downlink systems exist [24]–[26]. Reference [24] proposes a VLSI design for vector-perturbation precoding in small-scale MIMO systems with infinite-precision DACs. The papers [25] and [26] discuss hardware implementations for approximate linear and ZF/MRC-based precoding, respectively, for massive MU-MIMO systems with infinite-precision DACs. Unfortunately, both of these publications do not provide detailed FPGA implementation results. Hence, to the best of our knowledge, the VLSI designs proposed in this paper are the first hardware implementations reported in the open literature that are suitable for precoding in 1-bit massive MU-MIMO systems.

D. Notation

Lowercase and uppercase boldface letters designate column vectors and matrices, respectively. For a matrix $\mathbf{A}$, we denote its transpose, Hermitian transpose, complex conjugate, and matrix ℓ_2-norm by $\mathbf{A}^T$, $\mathbf{A}^H$, $\mathbf{A}^*$, and $\|\mathbf{A}\|_{2,2}$, respectively; the entry on the kth row and on the ℓth column of $\mathbf{A}$ is $[\mathbf{A}]_{k,\ell}$. The $M \times M$ identity matrix is denoted by $\mathbf{I}_M$ and the $M \times N$ all-zeros matrix is denoted by $\mathbf{0}_{M\times N}$. For a vector $\mathbf{a}$, the kth entry is $[\mathbf{a}]_k$ and we use $\|\mathbf{a}\|_2$ to denote the ℓ_2-norm of the vector $\mathbf{a}$. The real and imaginary parts of a complex vector $\mathbf{a}$ are $\Re\{\mathbf{a}\}$ and $\Im\{\mathbf{a}\}$, respectively. The signum function $\mathrm{sgn}(\cdot)$ is defined as $\mathrm{sgn}(a) = +1$ for $a \geq 0$ and $\mathrm{sgn}(a) = -1$ for $a < 0$ and is applied entry-wise to vectors. The multivariate complex-valued circularly-symmetric Gaussian probability density function (PDF) with covariance matrix $\mathbf{K}$ is denoted by $\mathcal{CN}(\mathbf{0}, \mathbf{K})$. We use $\mathbb{E}_{\mathbf{x}}[\cdot]$ to denote expectation with respect to the random vector $\mathbf{x}$.

E. Paper Outline

The rest of this paper is organized as follows. In Section II, we introduce the system model and formulate the precoding problem for systems with 1-bit DACs. In Section III, we propose two new 1-bit precoding algorithms, namely C1PO and C2PO. In Section IV and Section V, we detail our VLSI architectures for C1PO and C2PO, respectively. In Section VI, we show numerical simulations, reference FPGA implementation results, and a comparison with an MRC-based baseline precoder. We conclude the paper in Section VII. All proofs are relegated to Appendices A and B.

II. System Model and 1-bit Precoding

We start by introducing the downlink system model and then provide the necessary details about optimal precoding in 1-bit massive MU-MIMO systems.

A. Downlink System Model

We focus on the downlink of a single-cell, narrowband massive MU-MIMO system as illustrated in Figure 1. The

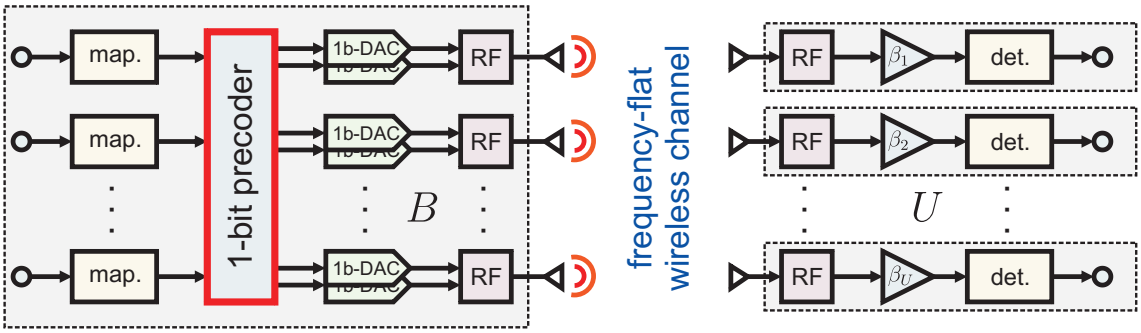


Fig. 1. Overview of a massive MU-MIMO downlink system with 1-bit DACs. Left: B antenna massive MU-MIMO BS containing a 1-bit precoder that mitigates multi-user interference and quantization artifacts in the 1-bit DACs; Right: U single-antenna UEs.

system consists of a B-antenna BS that serves $U \leq B$ single-antenna[1] UEs simultaneously and in the same frequency band. We use the standard input-output relation $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}$ to model the narrowband downlink channel [2]. Here, the vector $\mathbf{y} = [y_1, \ldots, y_U]^T$ contains the received signals at all UEs, where $y_u \in \mathbb{C}$ is the signal received at the uth UE. The matrix $\mathbf{H} \in \mathbb{C}^{U \times B}$ represents the downlink channel. The so-called *precoded vector* is denoted by $\mathbf{x} \in \mathcal{X}^B$, where $\mathcal{X}$ represents the transmit alphabet; this set coincides with the set $\mathbb{C}$ of complex numbers in the case of infinite-precision DACs. In 1-bit massive MU-MIMO systems, the in-phase and quadrature components are quantized separately using a pair of 1-bit DACs running at Nyquist rate and hence, the per-antenna quaternary transmit alphabet is $\mathcal{X} = \{\pm\ell \pm j\ell\}$ for a given (and fixed) $\ell > 0$ that determines the transmit power. The vector $\mathbf{n} \in \mathbb{C}^U$ models i.i.d. circularly-symmetric complex Gaussian noise with variance N_0 per complex entry, i.e., $n_u \sim \mathcal{CN}(0, N_0)$, for $u = 1, \ldots, U$. In what follows, we assume that the realization of the channel matrix $\mathbf{H}$ and the noise variance N_0 are perfectly known at the BS.[2]

B. Precoding Basics

The main purpose of precoding is to transmit constellation points $s_u \in \mathcal{O}$ to each UE $u = 1, \ldots, U$, where $\mathcal{O}$ is the constellation set (e.g., 16-QAM). The BS uses the available channel state information (CSI) to precode the symbol vector $\mathbf{s} = [s_1, \ldots, s_U]^T$ into the precoded vector $\mathbf{x} \in \mathcal{X}^B$. Throughout the paper, we assume that the precoded vector $\mathbf{x}$ must satisfy an instantaneous power constraint $\|\mathbf{x}\|_2^2 = P$; this leads to $\mathcal{X} = \{\pm\ell \pm j\ell\}$ with $\ell = \sqrt{P/(2B)}$.

Coherent transmission of data using multiple BS antennas leads to an array gain, which depends on the realization of the fading channel and the precoding method. As in [6], [7], we assume that the uth UE is able to rescale its received signals y_u by a factor[3] $\beta_u \in \mathbb{C}$ in order to compute an estimate $\hat{s}_u = \beta_u y_u$ for $u = 1, \ldots, U$ of the transmitted symbol $s_u \in \mathcal{O}$.

Since the UEs cannot perform joint processing to recover the transmitted data, precoding must simultaneously reduce MUI and increase signal power at all UEs [27]. To accomplish these goals, there exist multiple formulations of this optimization problem based on different performance metrics, e.g., sum-rate throughput or error rate (see [28] for a survey). As in [6], [7], we will focus exclusively on precoders that minimize the mean-square error (MSE) between the estimated symbol vector $\hat{\mathbf{s}} = \left[\hat{s}_1, \ldots, \hat{s}_U\right]^T = \beta\mathbf{y}$ and the transmitted symbol vector $\mathbf{s}$ given by

$$\mathbb{E}_{\mathbf{n}}\left[\|\mathbf{s} - \hat{\mathbf{s}}\|_2^2\right] = \|\mathbf{s} - \beta\mathbf{H}\mathbf{x}\|_2^2 + |\beta|^2 U N_0, \qquad (1)$$

where we restrict ourselves to the case in which the precoder results in the same *precoding factor* β for all UEs. Hence, in the remainder of this paper we shall assume that $\beta_u = \beta$ for $u = 1, \ldots, U$. In [7] it is shown that the UEs are able to accurately estimate the precoding factor β using pilot-based transmission in block-fading scenarios.

In the infinite-precision case, an MSE-optimal linear precoder multiplies the symbol vector $\mathbf{s}$ with a precoding matrix $\mathbf{P} \in \mathbb{C}^{B \times U}$ so that (1) is minimized on average over all possible transmit vectors $\mathbf{s}$ subject to the power constraint. This problem, which has been studied extensively for the case of infinite-precision DACs [29], [30], enables the design of low-complexity linear precoding algorithms [2].

C. MSE-Optimal 1-bit Precoding Problem

In the 1-bit case, linear-quantized precoders perform first linear precoding and then quantize the result to the finite transmit set $\mathcal{X}^B$ as

$$\mathbf{x} = \sqrt{\frac{P}{2B}}\left(\operatorname{sgn}(\Re\{\mathbf{P}\mathbf{s}\}) + j\operatorname{sgn}(\Im\{\mathbf{P}\mathbf{s}\})\right)$$

for a given precoding matrix $\mathbf{P}$. Linear-quantized precoders can be analyzed theoretically and typically exhibit low complexity [6]. However, as recently shown in [1], [6]–[9], significant performance improvements can be obtained by using sophisticated nonlinear precoding methods.

[1]For simplicity, we focus on single-antenna UEs; the model can easily be expanded to support multi-antenna UEs.

[2]Knowledge of $\mathbf{H}$ is typically acquired via training in the uplink in a time-division duplexing system [2]. As discussed in [6], channel estimation errors incur only a small performance loss. Knowledge of the noise variance N_0 at the BS can be obtained by explicit feedback from the UEs to the BS.

[3]In contrast to references [6], [7], which assumed a real-valued factors β_u, $u = 1, \ldots, U$, we allow these factors to be complex-valued.

One way to design such nonlinear precoders is to solve the following MSE-optimal 1-bit precoding problem (OPP), which simultaneously finds the optimal precoding vector $\mathbf{x}^{\text{OPP}}$ and the associated precoding factor β^{OPP}:

$$\text{(OPP)} \qquad \underset{\mathbf{x}\in\mathcal{X}^B,\,\beta\in\mathbb{C}}{\text{minimize}} \ \|\mathbf{s}-\beta\mathbf{H}\mathbf{x}\|_2^2+|\beta|^2UN_0. \tag{2}$$

We emphasize that for a fixed value of β, the problem (OPP) is a closest vector problem that is known to be NP-hard [31]–[33]; this implies that there exists no known algorithm to solve it efficiently for large values of B. In [6], [7], approximate methods for solving (2) using convex relaxation have been proposed, such as the squared-infinity norm Douglas-Rachford splitting (SQUID) algorithm. Such relaxation-based methods, however, still require high computational complexity, which prevents their deployment in practical systems.

III. 1-bit Precoding via Biconvex Relaxation

Since the problem (OPP) is of combinatorial nature, a brute-force search for a solution is intractable in massive MU-MIMO systems with hundreds of BS antennas. We next propose two nonlinear precoding algorithms that yield approximate but accurate solutions at low computational complexity.

A. Approximating (OPP)

To solve (OPP) efficiently, we use the BCR framework put forward in [14], which was initially proposed for solving large semidefinite programs that appear in computer vision. In order to use this framework, we first simplify the objective function of (OPP) by assuming that $N_0\to 0$, i.e., we assume that the system operates in the high-SNR regime. As in [1, Eq. 3], we take a leap of faith with the approximation

$$\min_{\mathbf{x}\in\mathcal{X}^B}\min_{\beta\in\mathbb{C}}\|\mathbf{s}-\beta\mathbf{H}\mathbf{x}\|_2^2\approx\min_{\mathbf{x}\in\mathcal{X}^B}\min_{\alpha\in\mathbb{C}}\|\alpha\mathbf{s}-\mathbf{H}\mathbf{x}\|_2^2. \tag{3}$$

This approximation can be justified by noting that if we can find a precoded vector $\mathbf{x}\in\mathcal{X}^B$ for which $\mathbf{s}=\mathbf{H}\mathbf{x}$, then both problems in (3) are indeed equivalent. These approximations allow us to rewrite (OPP) as follows:

$$\text{(OPP}^*\text{)} \qquad \hat{\mathbf{x}}=\underset{\mathbf{x}\in\mathcal{X}^B,\,\alpha\in\mathbb{C}}{\arg\min}\ \|\alpha\mathbf{s}-\mathbf{H}\mathbf{x}\|_2^2.$$

We next get rid of the parameter α in (OPP*). For a fixed $\mathbf{x}$, the optimal parameter $\hat{\alpha}(\mathbf{x})$ that minimizes the objective function of (OPP*) is given by

$$\hat{\alpha}(\mathbf{x})=\underset{\alpha\in\mathbb{C}}{\arg\min}\,\|\alpha\mathbf{s}-\mathbf{H}\mathbf{x}\|_2^2=\frac{\mathbf{s}^H\mathbf{H}\mathbf{x}}{\|\mathbf{s}\|_2^2}.$$

By inserting $\hat{\alpha}(\mathbf{x})$ into the objective function of (OPP*), we obtain

$$\|\hat{\alpha}(\mathbf{x})\mathbf{s}-\mathbf{H}\mathbf{x}\|_2^2=\|\mathbf{A}\mathbf{x}\|_2^2 \tag{4}$$

with

$$\mathbf{A}=\mathbf{Q}\mathbf{H}\quad\text{and}\quad\mathbf{Q}=\mathbf{I}_U-\frac{\mathbf{s}\mathbf{s}^H}{\|\mathbf{s}\|_2^2}, \tag{5}$$

where the matrix $\mathbf{Q}\in\mathbb{C}^{U\times U}$ is a projection onto the orthogonal complement of the space spanned by the symbol vector $\mathbf{s}$. Using (4), the problem (OPP*) can be simplified to

$$\text{(OPP}^*\text{)} \qquad \hat{\mathbf{x}}=\underset{\mathbf{x}\in\mathcal{X}^B}{\arg\min}\ \|\mathbf{A}\mathbf{x}\|_2^2,$$

which remains to be a closest vector problem. Nevertheless, the specific form of (OPP*) enables us to use BCR to efficiently compute approximate but accurate solutions.

B. Biconvex Relaxation (BCR)

To solve (OPP*) using BCR, we first introduce a copy $\mathbf{z}$ of the vector $\mathbf{x}$, and replace (OPP*) with the approximation

$$\hat{\mathbf{x}}=\underset{\mathbf{x}\in\mathcal{X}^B,\,\mathbf{z}\in\mathbb{C}^B}{\arg\min}\ \|\mathbf{A}\mathbf{z}\|_2^2+\gamma\|\mathbf{z}-\mathbf{x}\|_2^2,$$

where $\gamma>0$ is a (fixed) regularization parameter. We next relax the nonconvex alphabet constraint $\mathbf{x}\in\mathcal{X}^B$ to its convex envelope given by

$$\mathcal{B}^B=\Bigg\{\mathbf{c}\in\mathbb{C}^B\ \Bigg|\ |\Re\{c_b\}|\le\sqrt{\frac{P}{2B}},\ |\Im\{c_b\}|\le\sqrt{\frac{P}{2B}},\ b=1,\dots,B\Bigg\}, \tag{6}$$

which allows us to convexify the precoding problem as follows:

$$\hat{\mathbf{x}}=\underset{\mathbf{x}\in\mathcal{B}^B,\,\mathbf{z}\in\mathbb{C}^B}{\arg\min}\ \|\mathbf{A}\mathbf{z}\|_2^2+\gamma\|\mathbf{z}-\mathbf{x}\|_2^2.$$

Unfortunately, solving this problem yields, in general, the all-zeros vector, i.e., $\mathbf{x}=\mathbf{0}_{B\times 1}$. One of the key ideas of BCR is to force the solution of this new problem to satisfy the constraints in (6) with equality. This can be accomplished by including a nonconvex regularization term in the objective that promotes large values of $\mathbf{x}$. As suggested in [14], we use a negative ℓ_2-norm term to obtain the following biconvex relaxation (BCR) optimization problem:

$$\text{(BCR)} \quad \hat{\mathbf{x}}^{\text{BCR}}=\underset{\mathbf{x}\in\mathcal{B}^B,\,\mathbf{z}\in\mathbb{C}^B}{\arg\min}\ \|\mathbf{A}\mathbf{z}\|_2^2+\gamma\|\mathbf{z}-\mathbf{x}\|_2^2-\delta\|\mathbf{x}\|_2^2, \tag{7}$$

where $\delta>0$ is a (fixed) regularization parameter. When $\delta<\gamma$, the formulation (7) is *bi-convex* (i.e., the minimization with respect to $\mathbf{x}$ is convex when $\mathbf{z}$ is fixed, and vice versa). Robust parameter choices are $\gamma=\|\mathbf{A}^H\mathbf{A}\|_{2,2}$ and $\gamma/\delta=2$; see [14] for more details. In practice, we tune the parameters γ and δ to further improve the empirical performance of our algorithms.

C. C1PO: biConvex 1-bit PrecOding

We noted above that the BCR problem is biconvex, meaning that minimization with respect to $\mathbf{x}$ alone (with $\mathbf{z}$ fixed) or $\mathbf{z}$ alone (with $\mathbf{x}$ fixed) is convex. Hence, as done in [14], we can solve the BCR problem approximately using alternating minimization. Since the problem is nonconvex, initialization critically affects the performance of our algorithm. We initialize our algorithm with the MRC precoded vector $\mathbf{x}^{(1)}=\mathbf{H}^H\mathbf{s}$, which yields excellent performance in practice and can be computed efficiently. Then, we solve for $\mathbf{z}$ while holding $\mathbf{x}$

fixed; afterwards, we solve for $\mathbf{x}$ while holding $\mathbf{z}$ fixed. Specifically, we repeat the following procedure

$$\mathbf{z}^{(t+1)} = \arg\min_{\mathbf{z}\in\mathbb{C}^B} \|\mathbf{A}\mathbf{z}\|_2^2 + \gamma\|\mathbf{z}-\mathbf{x}^{(t)}\|_2^2$$
$$\mathbf{x}^{(t+1)} = \arg\min_{\mathbf{x}\in\mathcal{B}^B} \gamma\|\mathbf{z}^{(t+1)}-\mathbf{x}\|_2^2 - \delta\|\mathbf{x}\|_2^2$$

for $t = 1, 2, \ldots, t_{\max}$, where $t_{\max}$ is the maximum number of iterations. Both steps are convex optimization problems that can be solved efficiently in closed form. Hence, the above iterative procedure reduces to the following simple algorithm, which we call C1PO (short for biConvex 1-bit PrecOding).

Algorithm 1 (C1PO)**.** *Set* $\mathbf{A}$ *as in (5), initialize* $\mathbf{x}^{(1)} = \mathbf{H}^H\mathbf{s}$*, and fix the parameters* δ *and* γ *so that* $0 < \delta < \gamma$*. Then, for every iteration* $t = 1, 2, \ldots, t_{max}$*, compute*

$$\mathbf{z}^{(t+1)} = (\mathbf{I}_B + \gamma^{-1}\mathbf{A}^H\mathbf{A})^{-1}\mathbf{x}^{(t)} \tag{8}$$
$$\mathbf{x}^{(t+1)} = \mathrm{proj}(\mathbf{z}^{(t+1)}). \tag{9}$$

Here, the expansion-reprojection operator $\mathrm{proj}(\cdot)$ *is*

$$\mathrm{proj}(z) = \mathrm{sgn}(\Re\{z\})\min\left\{\frac{\gamma}{\gamma-\delta}|\Re\{z\}|, \sqrt{\frac{P}{2B}}\right\} + j\,\mathrm{sgn}(\Im\{z\})\min\left\{\frac{\gamma}{\gamma-\delta}|\Im\{z\}|, \sqrt{\frac{P}{2B}}\right\}$$

and is applied element-wise to the vector $\mathbf{z}^{(t+1)}$*. In the last iteration* t_{max}*, the output* $\mathbf{x}^{(t_{max}+1)}$ *of C1PO is quantized to the quaternary alphabet* $\mathcal{X} = \{\pm\ell \pm j\ell\}$ *with* $\ell = \sqrt{P/(2B)}$ *as follows:*

$$\hat{\mathbf{x}} = \sqrt{\frac{P}{2B}}\left(\mathrm{sgn}\left(\Re\left\{\mathbf{x}^{(t_{max}+1)}\right\}\right) + j\,\mathrm{sgn}\left(\Im\left\{\mathbf{x}^{(t_{max}+1)}\right\}\right)\right). \tag{10}$$

Because C1PO decreases the objective function (7) on every variable update, and the objective is bounded from below, the objective values corresponding to the iterates $\{\mathbf{x}^{(t)}, \mathbf{z}^{(t)}\}$ form a convergent sequence. However, by exploiting the biconvex structure of our problem, we can prove the following stronger result; the proof is given in Appendix A.

Theorem 1. *Any limit point of the sequence* $\{\mathbf{x}^{(t)}, \mathbf{z}^{(t)}\}$ *generated by C1PO is a stationary point of* (7)*.*

The main computations performed by C1PO in Algorithm 1 are (i) the $B \times B$ matrix inversion $\mathbf{G} = (\mathbf{I}_B + \gamma^{-1}\mathbf{A}^H\mathbf{A})^{-1}$, which can be computed once during a preprocessing stage, and (ii) the per-iteration matrix-vector multiplication $\mathbf{z}^{(t+1)} = \mathbf{G}\mathbf{x}^{(t)}$ in step (8); the complexity of the projection in step (9) is negligible. Unfortunately, the complexity of the matrix inversion, evaluated in terms of operations,[4] scales roughly with B^3 and the complexity of the per-iteration matrix-vector product with B^2. Both of these tasks are particularly inefficient for massive MU-MIMO systems with a large number of BS antennas. Therefore, we next propose an algorithmic variant that avoids both of these issues and whose complexity scales more favorably with the number of BS antennas.

D. Fast Algorithm for Very-Large Systems: C2PO

To obtain our alternative algorithm, we start from the BCR formulation in (7) but rather than introducing the auxiliary variable $\mathbf{z}$, we attempt to solve the following nonconvex problem:[5]

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}\in\mathcal{B}^B} \ \frac{1}{2}\|\mathbf{A}\mathbf{x}\|_2^2 - \frac{\delta}{2}\|\mathbf{x}\|_2^2. \tag{11}$$

We use forward-backward splitting (FBS) [34]–[36], a computationally efficient method to solve large convex problems. Since the problem in (11) is nonconvex, FBS is not guaranteed to converge to the optimal solution. Nevertheless, as shown in Section VI, the proposed algorithm performs well in practice.

FBS is an efficient iterative procedure to solve convex optimization problems of the form

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} f(\mathbf{x}) + g(\mathbf{x}),$$

where the function f is smooth and convex, and the function g is convex but not necessarily smooth or bounded. FBS consists of the following iteration [34], [35]:

$$\mathbf{z}^{(t+1)} = \mathbf{x}^{(t)} - \tau^{(t)}\nabla f(\mathbf{x}^{(t)})$$
$$\mathbf{x}^{(t+1)} = \mathrm{prox}_g\left(\mathbf{z}^{(t+1)}; \tau^{(t)}\right)$$

for $t = 1, 2, \ldots, t_{\max}$ or until convergence. Here, $\nabla f(\mathbf{x})$ is the gradient of the smooth function f, and the so-called proximal operator for the function g is defined as follows [37]:

$$\mathrm{prox}_g(\mathbf{z};\tau) = \arg\min_{\mathbf{x}}\left\{\tau g(\mathbf{z}) + \frac{1}{2}\|\mathbf{x}-\mathbf{z}\|_2^2\right\}.$$

The sequence $\{\tau^{(t)} > 0\}$ contains suitably chosen step-size parameters. For the problem (11), we show below that FBS monotonically decreases the objective (11) for any constant step size that satisfies $\tau^{(t)} = \tau < \|\mathbf{A}^H\mathbf{A}\|_{2,2}^{-1}$.

In order to approximately solve (11) using FBS, we set

$$f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x}\|_2^2 \ \text{ and } \ g(\mathbf{x}) = \chi(\mathbf{x}\in\mathcal{B}^B) - \frac{\delta}{2}\|\mathbf{x}\|_2^2, \tag{12}$$

where χ is a *characteristic function* that is zero if the condition $\mathbf{x} \in \mathcal{B}^B$ is met and infinity otherwise. For this choice of f and g, the gradient is given by $\nabla f(\mathbf{x}) = \mathbf{A}^H\mathbf{A}\mathbf{x}$ and the proximal operator is given by the expansion-reprojection operation

$$\mathrm{prox}_g(z) = \mathrm{sgn}(\Re\{z\})\min\left\{\frac{1}{1-\tau\delta}|\Re\{z\}|, \sqrt{\frac{P}{2B}}\right\} + j\,\mathrm{sgn}(\Im\{z\})\min\left\{\frac{1}{1-\tau\delta}|\Im\{z\}|, \sqrt{\frac{P}{2B}}\right\}, \tag{13}$$

[4]For simplicity, we count the number of complex-valued multiplications to characterize the operation count.

[5]To simplify notation, we have divided both terms in the objective function by a factor of two; this scaling does not affect the result.

which is valid for $\tau\delta < 1$ and applied element-wise to vectors. By using FBS with the above-mentioned ingredients, we obtain the following simple algorithm, which we call C2PO.

Algorithm 2 (C2PO)**.** *Set* $\mathbf{A}$ *as in (5). Initialize* $\mathbf{x}^{(1)} = \mathbf{H}^H\mathbf{s}$ *and fix the parameters* δ *and* τ *so that* $\tau\delta < 1$*. Then, for every iteration* $t = 1, 2, \ldots, t_{max}$ *compute:*

$$\mathbf{z}^{(t+1)} = \mathbf{x}^{(t)} - \tau\mathbf{A}^H\mathbf{A}\mathbf{x}^{(t)} \tag{14}$$

$$\mathbf{x}^{(t+1)} = \text{prox}_g(\overline{\mathbf{z}}^{(t+1)}; \tau). \tag{15}$$

Here, the prox_g *operator is given in (13) and is applied element-wise to the vector* $\mathbf{z}^{(t+1)}$*. In the last iteration* t_{max}*, the output* $\mathbf{x}^{(t_{max}+1)}$ *of C2PO is quantized to the quaternary alphabet* $\mathcal{X}$ *as in (10).*

The following result shows that C2PO is well behaved, provided that the step size is chosen appropriately; the proof is given in Appendix B.

Theorem 2. *Suppose the step size used in C2PO satisfies* $\tau < \|\mathbf{A}^H\mathbf{A}\|_{2,2}^{-1}$*, and* $\tau\delta < 1$*. Then, C2PO decreases the objective* (11) *monotonically, and any limit point of the iterates* $\{\mathbf{x}^{(t)}\}$ *is a stationary point.*

The most complex operation of C2PO (Algorithm 2) is the the matrix-vector multiplication in step (14). In contrast to C1PO (Algorithm 1), however, this step requires a minimal amount of preprocessing and can be computed efficiently, especially for large BS antenna arrays. To see this, we rewrite $\mathbf{A}^H\mathbf{A}$, where $\mathbf{A}$ was given in (5), as follows:

$$\mathbf{A}^H\mathbf{A} = \mathbf{H}^H\mathbf{H} - \frac{\mathbf{H}^H\mathbf{s}\mathbf{s}^H\mathbf{H}}{\|\mathbf{s}\|_2^2} = \mathbf{H}^H\mathbf{H} - \mathbf{v}\mathbf{v}^H = \overline{\mathbf{H}}^H\overline{\mathbf{H}}.$$

Here, $\mathbf{v} = \mathbf{H}^H\mathbf{s}/\|\mathbf{s}\|_2$ is a normalized version of the MRC vector and $\overline{\mathbf{H}} = [\mathbf{H}; j\mathbf{v}^H]$ is an augmented $(U+1)\times B$ matrix. With these definitions, we can now simplify step (14) to

$$\mathbf{z}^{(t+1)} = \mathbf{x}^{(t)} - \tau\overline{\mathbf{H}}^H\overline{\mathbf{H}}\mathbf{x}^{(t)}, \tag{16}$$

where we first compute $\mathbf{w} = \overline{\mathbf{H}}(\tau\mathbf{x}^{(t)})$, then $\mathbf{w}' = \overline{\mathbf{H}}^H\mathbf{w}$, and finally $\mathbf{z}^{(t+1)} = \mathbf{x}^{(t)} - \mathbf{w}'$. As a result, we conclude that each iteration of C2PO requires only two matrix-vector products with a cost of roughly $2B(U+1)$ operations (in contrast to B^2 operations for C1PO). In addition, the preprocessing stage of this algorithm only needs to compute the normalized MRC vector $\mathbf{v}$, which requires roughly BU operations (in contrast to B^3 operations for C1PO). Hence, the complexity of C2PO can be significantly lower than that of C1PO, especially since the antenna configurations of typical massive MU-MIMO systems satisfy $U \ll B$. As we will show in Section VI, the hardware efficiency of C2PO is superior to that of C1PO for large BS antenna arrays and the error-rate performance is comparable.

E. Alternative Derivation of C2PO

It is interesting to note that there is a strong connection between Algorithm 1 and Algorithm 2. In fact, one can obtain C2PO directly from C1PO using the following well-known series expansion. Let $\|\mathbf{A}^H\mathbf{A}\|_{2,2} < \gamma$. Then, we have the following Neumann series expansion [38]:

$$(\mathbf{I}_B + \gamma^{-1}\mathbf{A}^H\mathbf{A})^{-1} = \sum_{n=1}^{\infty}(-\gamma^{-1}\mathbf{A}^H\mathbf{A})^n.$$

As suggested in [20], we can approximate the inverse by truncating the series to the two first terms:

$$(\mathbf{I}_B + \gamma^{-1}\mathbf{A}^H\mathbf{A})^{-1} \approx \mathbf{I}_B - \gamma^{-1}\mathbf{A}^H\mathbf{A}.$$

By using this approximation in step (8) of Algorithm 1, we immediately obtain Algorithm 2 after setting $\gamma^{-1} = \tau$. Note that the Neumann series expansion is only convergent for $\|\mathbf{A}^H\mathbf{A}\|_{2,2} < \gamma$, which corresponds to the step size restriction $\tau < \|\mathbf{A}^T\mathbf{A}\|_{2,2}^{-1}$; this is the step size requirement in Theorem 2.

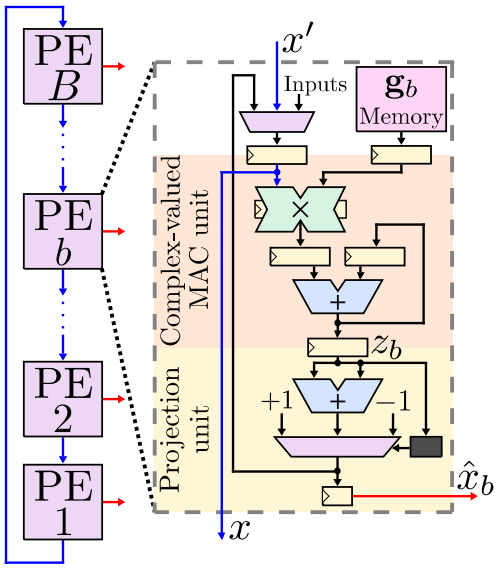


Fig. 2. High-level block diagram of the VLSI architecture for C1PO. We use a linear array of B processing elements (PEs) that enables us to achieve high throughput at low hardware complexity.

IV. VLSI Design for C1PO

We now present a high-throughput VLSI architecture for C1PO as in Algorithm 1. We then discuss the key optimizations performed in our FPGA implementation.

A. Architecture Overview

The proposed VLSI architecture that implements C1PO as detailed in Algorithm 1 is shown in Figure 2. Our architecture consists of a linear array of B identical *processing elements (PEs)* that share a common control unit. The PEs essentially compute the complex-valued matrix-vector product in (8) in a sequential, column-by-column manner followed by the projection operation in (9). Each PE $b = 1, 2, \ldots, B$ consists of three main building blocks: (i) a $\mathbf{g}_b$-memory, (ii) a complex-valued multiply-accumulate (MAC) unit, and (iii) a projection unit. For the bth PE, the $\mathbf{g}_b$-memory stores the bth row of the matrix $\mathbf{G} = (\mathbf{I}_B + \gamma^{-1}\mathbf{A}^H\mathbf{A})^{-1}$, which we assume was computed during a preprocessing stage. The complex-valued MAC unit is used by each PE to sequentially compute an entry of the output

vector $\mathbf{z}^{(t+1)}$ on line (8), while the entries of the vector $\mathbf{x}^{(t)}$ are exchanged between the PEs in a cyclic fashion; this is done to avoid an architecture with a centralized $\mathbf{x}^{(t)}$ memory that would suffer from a high fan-out because the memory's output has to be distributed to all the PEs. The projection unit implements the expansion-reprojection operator $\mathrm{proj}(\cdot)$ on line (9) in a hardware-friendly manner. The outputs of the projection unit are also used to generate the quaternary outputs of the 1-bit precoder; to this end, each PE simply takes the sign bits of the complex-valued output vector $\mathbf{x}^{(t+1)}$.

B. Architecture Operation

We now detail the (rather technical) operation of the C1PO architecture illustrated in Figure 2. In the first iteration (i.e., at $t = 1$), each PE b is initialized with the bth entry of the vector $\mathbf{x}^{(1)}$. Furthermore, the entries of the $\mathbf{g}_b$-memory are stored so that the first memory address corresponds to $[\mathbf{G}]_{b,b}$, the second address to $[\mathbf{G}]_{b,b+1}$, and so forth (addresses wrap around).

In the first clock cycle, each PE b computes $[\mathbf{G}]_{b,b}[\mathbf{x}^{(t)}]_b$ and the result is stored in the accumulator. As shown on the left side of Figure 2, in the same clock cycle the bth PE passes the value $[\mathbf{x}^{(t)}]_b$ to PE $(b-1)$, while it receives the value $[\mathbf{x}^{(t)}]_{b+1}$ from PE $(b+1)$; PE 1 passes its value to PE B. In the second clock cycle, since the exchange operation made the element $[\mathbf{x}^{(t)}]_{b+1}$ available at PE b, each PE computes $[\mathbf{G}]_{b,b+1} \cdot [\mathbf{x}^{(t)}]_{b+1}$ and uses the accumulator to add it to the result of the previous cycle. Once again, in the same clock cycle, the bth PE passes the $\mathbf{x}^{(t)}$ entry that is currently being multiplied on its MAC unit to PE $(b-1)$; PE 1 passes its value to PE B. Consequently, in the third clock cycle, the bth PE will use the values $[\mathbf{G}]_{b,b+2}$ and $[\mathbf{x}^{(t)}]_{b+2}$ to continue performing MAC operations. By repeating this procedure B times, each entry of the vector $[\mathbf{x}^{(t)}]$ circulates through all the PEs exactly once, enabling each PE $b = 1, 2, \ldots, B$ to compute $[\mathbf{z}^{(t+1)}]_b$. Thus, the matrix-vector multiplication on line (8) is completed. Since the complex-valued MAC unit contains three pipeline stages, two clock cycles are required to flush the pipeline. Hence, the matrix-vector operation requires a latency of $B+2$ clock cycles. After $B+2$ clock cycles, the vector $\mathbf{z}^{(t+1)}$ is available at the outputs of the MAC units.

In the subsequent clock cycle, each PE projects their respective entry of the vector $\mathbf{z}^{(t+1)}$. The choice $\gamma/\delta = 5$, which implies that $\gamma/(\gamma-\delta) = 1.25$, works well for the considered antenna configurations. Furthermore, to reduce the hardware complexity, we assume $P = 2B$ so that the clipping threshold of the expansion-reprojection operator $\mathrm{proj}(\cdot)$ is 1. As a result, the $\mathrm{proj}(\cdot)$ operator in (9) is implemented by applying the following operations independently to the real and imaginary parts of $[\mathbf{z}^{(t+1)}]_b$: We multiply the real (or imaginary) part of $[\mathbf{z}^{(t+1)}]_b$ by 1.25; this is accomplished by adding the $[\mathbf{z}^{(t+1)}]_b$ value with a $2\times$ right-shifted version of itself. At the same time, the real (or imaginary) part of $[\mathbf{z}^{(t+1)}]_b$ is compared to -0.8 and $+0.8$. If the real (or imaginary) part of $[\mathbf{z}^{(t+1)}]_b$ is between these two numbers, then the projection unit outputs $1.25 \cdot [\mathbf{z}^{(t+1)}]_b$. If it is smaller than -0.8, then the projection unit generates -1; if it is larger than $+0.8$, it generates $+1$. The result from this projection is stored as the next iterate $[\mathbf{x}^{(t+1)}]_b$ in the input register of the complex-valued MAC unit, which completes one C1PO iteration. Since the projection requires an additional clock cycle, one C1PO iteration is completed in exactly $B+3$ clock cycles.

C. FPGA Implementation Details

To minimize the FPGA implementation complexity of C1PO, we exclusively use fixed-point arithmetic; see Section VI-A for the fixed-point error-rate performance of C1PO. To represent the entries of the vector $\mathbf{x}^{(t)}$, we use 12-bit fixed-point values with 5 fraction bits. The entries of the $\mathbf{G}$ matrix are represented using 10 bits with 9 fraction bits, and we use FPGA look-up tables (LUTs) as a distributed RAM to store these values. The complex-valued MAC unit uses 18 bits with 11 fraction bits; the projection unit uses 15 bits with 8 fraction bits. In our C1PO design, all adders and multipliers do not saturate, but wrap around; number resizing uses truncation. All complex-valued multipliers consist of four real-valued multipliers and two adders; we use the built-in DSP48 units for these operations.

V. VLSI Design for C2PO

We now present a high-throughput VLSI architecture for C2PO as in Algorithm 2. We then discuss the key optimizations used in our FPGA implementation.

A. Architecture Overview

The proposed VLSI architecture that implements C2PO (Algorithm 2) is shown in Figure 3. In what follows, we assume that B is a multiple of U and $B \gg U$. Our architecture consists of B/U linear arrays; each array consists of $U+1$ PEs and a control unit. The architecture divides the operation in (16) into two separate matrix-vector products: (i) $\mathbf{w} = \overline{\mathbf{H}}(\tau\mathbf{x}^{(t)})$ and (ii) $\mathbf{w}' = \overline{\mathbf{H}}^H\mathbf{w}$; see also the discussion at the end of Section III-D. Note that for the first matrix-vector product, the matrix $\overline{\mathbf{H}}$ has more columns (B) than rows ($U+1$); for the second matrix-vector product, the matrix $\overline{\mathbf{H}}^H$ has more rows than columns. Therefore, we will refer to the first matrix-vector product as the *wide product*, while the second one will be identified as the *tall product*. The final subtraction required to compute $\mathbf{z}^{(t+1)} = \mathbf{x}^{(t)} - \mathbf{w}'$ in (16) is incorporated into the tall-product operation; see Section V-B for more details.

To perform both the wide and tall products in a single computation unit, the matrix $\overline{\mathbf{H}}$ is divided into B/U sub-matrices $\widetilde{\mathbf{H}}_w \in \mathbb{C}^{(U+1)\times U}$, $w = 1, 2, \ldots, B/U$, so that $\overline{\mathbf{H}} = \big[\widetilde{\mathbf{H}}_1, \widetilde{\mathbf{H}}_2, \ldots, \widetilde{\mathbf{H}}_{B/U}\big]$. Analogously, the vector $\mathbf{x}^{(t)}$ is divided into B/U sub-vectors $\tilde{\mathbf{x}}_w^{(t)} \in \mathbb{C}^U$, $w = 1, 2, \ldots, B/U$, so that $\mathbf{x}^{(t)} = \big[\tilde{\mathbf{x}}_1^{(t)}; \tilde{\mathbf{x}}_2^{(t)}; \ldots; \tilde{\mathbf{x}}_{B/U}^{(t)}\big]$. We next outline the architectural principle of the wide and tall products.

(i) Wide Product: Each linear array takes one sub-matrix $\widetilde{\mathbf{H}}_w$ and the associated sub-vector $\tilde{\mathbf{x}}_w^{(t)}$ as its inputs and computes their product in a sequential, column-by-column manner. This operation is analogous to that of the C1PO architecture (cf. Section IV-B) and within each linear array, the entries of the scaled

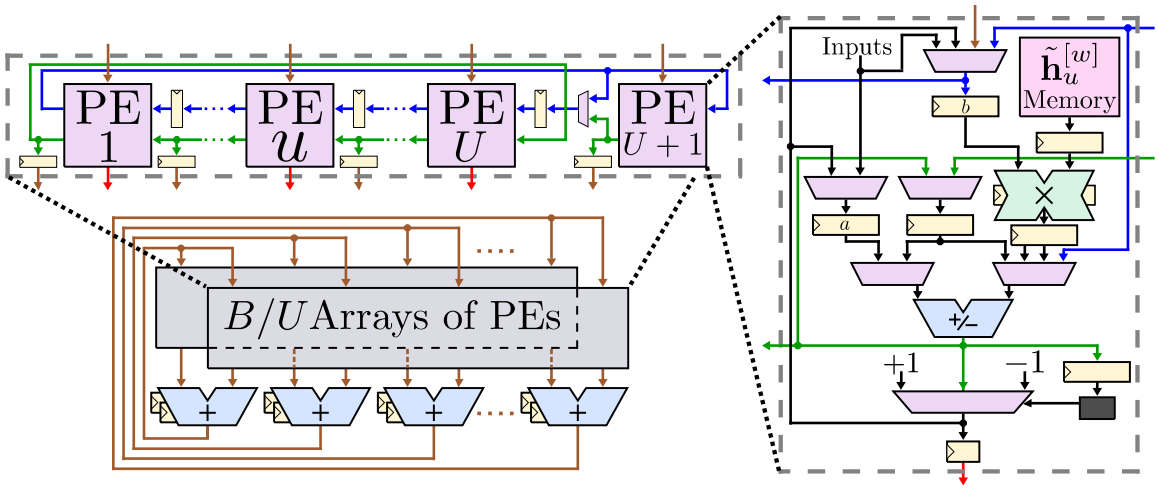


Fig. 3. High-level block diagram of the VLSI architecture for C2PO. We use B/U linear arrays, each consisting of $U+1$ processing elements (PEs), which enable us to achieve high throughput at low hardware complexity.

sub-vector $\tau\tilde{\mathbf{x}}_w^{(t)}$ are cyclically exchanged among the PEs. The resulting vectors $\mathbf{w}_w = \widetilde{\mathbf{H}}_w(\tau\tilde{\mathbf{x}}_w^{(t)})$ are then added to obtain

$$\mathbf{w} = \overline{\mathbf{H}}(\tau\mathbf{x}^{(t)}) = \sum_{w=1}^{B/U} \widetilde{\mathbf{H}}_w(\tau\tilde{\mathbf{x}}_w^{(t)}),$$

which completes the wide product. Each entry $[\mathbf{w}]_u$ of the resulting vector $\mathbf{w}$ is then stored in PE u of all linear arrays.

(ii) Tall Product: With the $\mathbf{w}$ vector available in all the linear arrays, each array now computes U entries of $\mathbf{z}^{(t+1)}$ by implementing $\mathbf{z}_w^{(t+1)} = \tilde{\mathbf{x}}_w^{(t)} - \widetilde{\mathbf{H}}_w^H\mathbf{w}$. Here, however, we use a sequential procedure in which the accumulated results are exchanged between PEs of the same array; see Section V-B for a detailed explanation. As a result, each linear array can project its computed $\mathbf{z}_w^{(t+1)}$ entries to generate the next iterate $\tilde{\mathbf{x}}_w^{(t+1)}$, which are then used by the same linear array to proceed with the next iteration. The sign bits of the new vector $\mathbf{x}^{(t+1)}$ correspond to the outputs of the C2PO architecture.

As in the C1PO architecture, each PE $u = 1, 2, \ldots, U+1$ is formed by three main units. The first unit is an $\tilde{\mathbf{h}}_u^{[w]}$-memory, which stores the uth row of the $\widetilde{\mathbf{H}}_w$ sub-matrix. The second unit is a complex-valued MAC unit, which supports (i) multiplications of $a \times b$ and $a \times b^*$, (ii) accumulation by addition or subtraction, and (iii) initialization of the accumulator with a non-zero value. The third unit is the projection unit, which is equivalent to the one of C1PO, although it is merged with the accumulator of the MAC unit.

B. Architecture Operation

We now provide the (rather technical) operation details of the C2PO architecture illustrated in Figure 3. Without loss of generality, we focus our description on the wth linear array of PEs, which operates on the $\widetilde{\mathbf{H}}_w$ sub-matrix and the $\tilde{\mathbf{x}}_w^{(t)}$ sub-vector. In the first iteration (i.e., at $t = 1$), the entry $[\tilde{\mathbf{x}}_w^{(1)}]_u$ and its scaled version $\tau[\tilde{\mathbf{x}}_w^{(1)}]_u$ are stored in PE $u = 1, 2, \ldots, U$ in two different registers: The value $[\tilde{\mathbf{x}}_w^{(1)}]_u$ is stored in the register labeled with "a" in Figure 3, which will later be used to initialize the accumulator in the complex-valued MAC unit; the value $\tau[\tilde{\mathbf{x}}_w^{(1)}]_u$ is stored at the input register of the MAC unit labeled with "b." We restrict the stepsize τ to be $2^{-\alpha}$ for some fixed $\alpha \in \mathbb{N}^+$, which enables us to acquire $\tau[\tilde{\mathbf{x}}_w^{(1)}]_u$ from a simple arithmetic right-shifted version of $[\tilde{\mathbf{x}}_w^{(1)}]_u$. The $(U+1)$th PE stores the same value $[\tilde{\mathbf{x}}_w^{(1)}]_1$ as that in PE 1. Similar to the C1PO architecture, the entries of the $\tilde{\mathbf{h}}_u^{[w]}$-memory are stored so that the first memory address contains $[\widetilde{\mathbf{H}}_w]_{u,u}$, the second address $[\widetilde{\mathbf{H}}_w]_{u,u+1}$, and so forth (addresses wrap around). For the $(U+1)$th PE, the first address of the $\tilde{\mathbf{h}}_{U+1}^{[w]}$-memory contains $[\widetilde{\mathbf{H}}_w]_{U+1,1}$, the second address contains $[\widetilde{\mathbf{H}}_w]_{U+1,2}$, etc.

(i) Wide Product: In the first clock cycle, each PE $u = 1, 2, \ldots, U$ computes $[\widetilde{\mathbf{H}}_w]_{u,u} \cdot [\tau\tilde{\mathbf{x}}_w^{(t)}]_u$ and stores the result in the accumulator. The $(U+1)$th PE computes $[\widetilde{\mathbf{H}}_w]_{U+1,1} \cdot [\tau\tilde{\mathbf{x}}_w^{(t)}]_1$. As shown in the upper left side of Figure 3, in the same clock cycle, the uth PE passes the value $[\tau\tilde{\mathbf{x}}_w^{(t)}]_u$ to PE $(u-1)$, while it receives the value $[\tau\tilde{\mathbf{x}}_w^{(t)}]_{u+1}$ from PE $(u+1)$; PE 1 passes its value to PE U, while PE $(U+1)$ does not pass anything. In the second clock cycle, since the cyclic exchange operation made the entry $[\tau\tilde{\mathbf{x}}_w^{(t)}]_{u+1}$ available at PE u, each PE computes $[\widetilde{\mathbf{H}}_w]_{u,u+1} \cdot [\tau\tilde{\mathbf{x}}_w^{(t)}]_{u+1}$ and uses the accumulator to add it to the result of the previous cycle. The $(U+1)$th PE uses the same value $\tau\tilde{\mathbf{x}}_w^{(t)}$ as PE 1; hence, it can compute $[\tau\tilde{\mathbf{x}}_w^{(t)}]_2 \cdot [\widetilde{\mathbf{H}}_w]_{U+1,2}$. Again, in the same clock cycle, the uth PE passes the $\tau\tilde{\mathbf{x}}_w^{(t)}$ entry that is currently being multiplied on its MAC unit to PE $(u-1)$; PE 1 passes its value to PE B, while PE $(U+1)$ does not pass anything. Consequently, in the third clock cycle, the uth PE will use the values $[\widetilde{\mathbf{H}}_w]_{u,u+2}$ and $[\tau\tilde{\mathbf{x}}_w^{(t)}]_{u+2}$ to continue performing MAC operations. During this third cycle, the $(U+1)$th PE will calculate the product $[\widetilde{\mathbf{H}}_w]_{U+1,3} \cdot [\tau\tilde{\mathbf{x}}_w^{(t)}]_3$. By repeating this procedure U times, each entry of the sub-vector $(\tau\tilde{\mathbf{x}}_w^{(t)})$ cycles through all the PEs exactly once, enabling the wth linear array of PEs to compute $\widetilde{\mathbf{H}}_w(\tau\tilde{\mathbf{x}}_w^{(t)})$. Since the complex-valued MAC unit contains three pipeline stages, two clock cycles are required to flush the pipeline. Hence, the previous matrix-vector operation has a latency of $U+2$ clock cycles. To complete the wide product, the vectors $\widetilde{\mathbf{H}}_w(\tau\tilde{\mathbf{x}}_w^{(t)})$ must be added. We use a

binary adder tree with $\log_2(B/U)$ pipeline stages. Hence, the vector $\mathbf{w}$ is computed after $U + \log_2(B/U) + 2$ clock cycles. The uth PE in each linear array stores the entry $[\mathbf{w}]_u$ in the MAC unit's input registered labeled with "b" in Figure 3.

(ii) Tall Product: In the next clock cycle, the computation of the tall-product starts. During the first clock cycle of the tall product computation, the PE $u = 1, 2, \ldots, U$ has available $[\mathbf{w}]_u$, as well as $[\widetilde{\mathbf{H}}_w]_{u,u}$, the first entry in its memory. The PE can then compute $[\widetilde{\mathbf{H}}_w]^*_{u,u} \cdot [\mathbf{w}]_u = [\widetilde{\mathbf{H}}^H_w]_{u,u} \cdot [\mathbf{w}]_u$. Using the accumulator, this product is then subtracted from $[\tilde{\mathbf{x}}^{(t)}_w]_u$, which was stored during the initialization phase in the register labeled with "a" in Figure 3. During the same clock cycle, the uth PE sends its accumulated result to the $(u-1)$th PE; PE 1 sends its accumulated result to PE U. Also, in the same clock cycle, the $(U+1)$th PE multiplies the conjugate of the first entry of its memory with its $\mathbf{w}$ entry. In words, the product $[\widetilde{\mathbf{H}}_w]^*_{U+1,1} \cdot [\mathbf{w}]_{U+1} = [\widetilde{\mathbf{H}}^H_w]_{1,U+1} \cdot [\mathbf{w}]_{U+1}$ is computed. The result is sent to the Uth PE. In the following clock cycles, this result will cycle through the linear array using the same wires and registers that were previously used to transfer the $\tau\tilde{\mathbf{x}}^{(t)}_w$ entries. In the second clock cycle, the $(u-1)$th PE multiplies the value $[\mathbf{w}]_{u-1}$ with $[\widetilde{\mathbf{H}}_w]^*_{u-1,u} = [\widetilde{\mathbf{H}}^H_w]_{u,u-1}$. The product is then subtracted from the accumulated value received from the uth PE during the previous cycle, and the new accumulated value is passed to the $(u-2)$th PE. In the same clock cycle, the $(U+1)$th PE multiplies the value $[\mathbf{w}]_{U+1}$ with $[\widetilde{\mathbf{H}}_w]^*_{U+1,2} = [\widetilde{\mathbf{H}}^H_w]_{2,U+1}$ and sends the result to the Uth PE, so it can cycle through the linear array. Furthermore, PE U passes the $[\widetilde{\mathbf{H}}^H_w]_{1,U+1} \cdot [\mathbf{w}]_{U+1}$ (previously received from the $(U+1)$th PE) to PE $(U-1)$. In the third clock cycle, the $(u-2)$th PE calculates $[\widetilde{\mathbf{H}}^H_w]_{u,u-2} \cdot [\mathbf{w}]_{u-2}$, subtracts it from the accumulated result received on the second cycle from the $(u-1)$th PE and passes the result to the $(u-3)$th PE. In the same clock cycle, the $(U+1)$th PE computes $[\widetilde{\mathbf{H}}^H_w]_{3,U+1} \cdot [\mathbf{w}]_{U+1}$ and sends it to PE U. Meanwhile, $[\widetilde{\mathbf{H}}^H_w]_{2,U+1} \cdot [\mathbf{w}]_{U+1}$ is passed from PE U to PE $(U-1)$ and $[\widetilde{\mathbf{H}}^H_w]_{1,U+1} \cdot [\mathbf{w}]_{U+1}$ is passed from PE $(U-1)$ to PE $(U-2)$. After repeating this procedure for U clock cycles, each PE $u = 1, 2, \ldots, U$ will contain the accumulated result for the uth entry of $\tilde{\mathbf{x}}^{(t)}_w - \widetilde{\mathbf{H}}^H_w \mathbf{w}$, received from the $(u+1)$th PE during the previous cycle. However, this accumulated result is missing the product $[\widetilde{\mathbf{H}}^H_w]_{u,U+1} \cdot [\mathbf{w}]_{U+1}$, which was computed and sent by the $(U+1)$th PE. Nonetheless, in the $(U+1)$th cycle of the tall product procedure, the uth PE receives the missing $[\widetilde{\mathbf{H}}^H_w]_{u,U+1} \cdot [\mathbf{w}]_{U+1}$ value from the $(u+1)$th PE. Thus, the $\mathbf{z}^{(t+1)} = \mathbf{x}^{(t)} - \overline{\mathbf{H}}^H \mathbf{w}$ entries are calculated after $U+1$ cycles. Since the complex MAC unit is used again, two additional clock cycles are required to flush its pipeline. Hence, $U+3$ cycles are used for the tall product.

Finally, in the subsequent clock cycle after the tall product is completed, the projection operator is applied in a similar fashion as for the C1PO architecture. The only difference is that now the accumulator of the MAC unit is used to multiply the real and imaginary parts of each $\mathbf{z}^{(t+1)}$ entry with 1.25, by adding each $\mathbf{z}^{(t+1)}$ entry with a $2\times$ right-shifted version of itself. The result from this projection is stored as the next iterate $[\tilde{\mathbf{x}}^{(t+1)}_w]_u$ in the two initialization locations previously mentioned, completing one C2PO iteration. Since the projection operation requires an additional clock cycle, a full C2PO iteration is completed in exactly $2U + \log_2(B/U) + 6$ clock cycles.

C. FPGA Implementation Details

As for the C1PO FPGA implementation, we exclusively use fixed-point arithmetic for the C2PO FPGA design; see Section VI-A for the fixed-point error-rate performance of C2PO. To represent the entries of the vector $\mathbf{x}^{(t)}$, we use 12-bit fixed-point values with 5 fraction bits. For the scaled $\tau\mathbf{x}^{(t)}$ values, we use 12-bit fixed-point values with 11 fraction bits. The entries of the $\overline{\mathbf{H}}$ matrix consist of 10 bits with 8 fraction bits, and we use FPGA look-up tables (LUTs) as a distributed RAM to store these values. The complex-valued MAC unit uses 18 bits with 15 fraction bits when doing the wide product and 11 fraction bits when doing the tall product; the projection unit uses 18 bits with 11 fraction bits. The adder tree uses 21 bits with 15 fraction bits. Identical to the C1PO implementation, all adders and multipliers do not saturate, but wrap around; number resizing uses truncation. All complex-valued multipliers are built with four real-valued multipliers and two adders; we use DSP48 units for these operations.

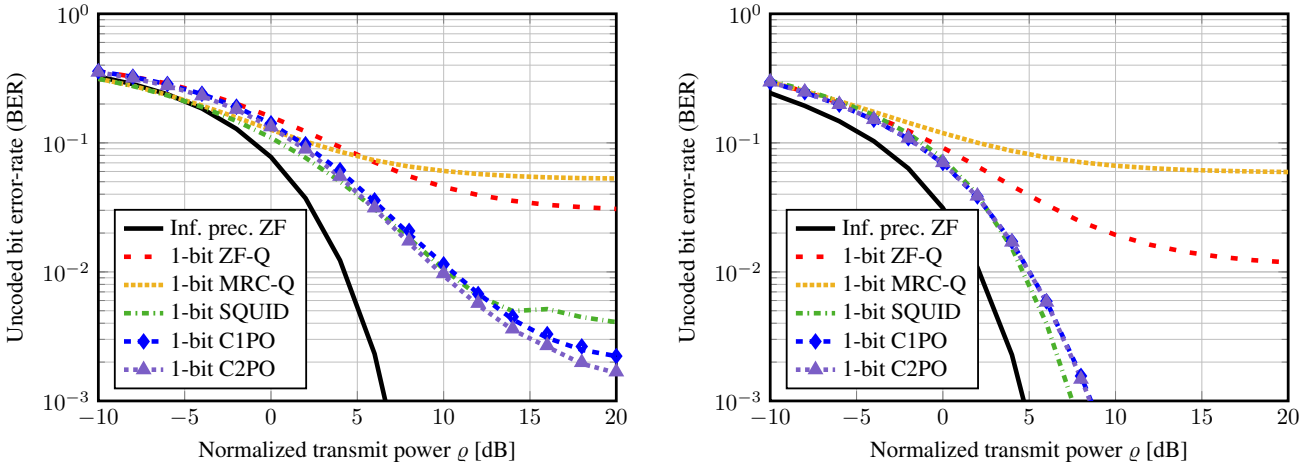
VI. Results

We now provide error-rate performance results for massive MU-MIMO systems and show reference FPGA implementation results for C1PO and C2PO.

A. Simulation Results

Figure 4 shows uncoded bit error rate (BER) curves versus the normalized transmit power $\varrho = P/N_0$ for massive MU-MIMO dowlink systems with $U = 16$ UEs for various precoding algorithms. In Figure 4(a) we consider the case of $B = 32$ BS antennas with BPSK, whereas in Figure 4(b) we consider the case of $B = 128$ BS antennas with 16-QAM. For both systems, we use Gray mapping, generate i.i.d. Rayleigh fading channel matrices, and average the BER over 10 000 Monte–Carlo trials. We compare ZF followed by quantization (ZF-Q), MRC followed by quantization (MRC-Q), the nonlinear SQUID algorithm proposed in [6], as well as C1PO and C2PO, for systems with 1-bit DACs. As a reference, we also include ZF with infinite-precision DACs (Inf. prec. ZF). SQUID runs $t_{\text{max}} = 50$ iterations; C1PO and C2PO both run $t_{\text{max}} = 24$ iterations. For all algorithms, the curves represent MATLAB floating-point performance; for C1PO and C2PO, the markers correspond to fixed-point performance of our hardware designs. Clearly, the fixed-point implementation loss of our hardware designs is negligible, i.e., less than 0.15 dB SNR at 1% uncoded BER for both considered scenarios.

For the 16×32 system (we use the notation $U \times B$ to refer to a downlink scenario with U users and B BS antennas) with BPSK, Figure 4(a) shows that all nonlinear precoders significantly outperform the linear-quantized precoders (ZF-Q and MRC-Q), which exhibit a high error floor. Furthermore, we see that C1PO and C2PO achieve similar performance as that of SQUID. At low-SNR, SQUID is marginally better, whereas C2PO achieves the best performance at high SNR, closely followed by C1PO and SQUID.



(a) $B = 32$, $U = 16$, and BPSK.

(b) $B = 256$, $U = 16$, and 16-QAM.

Fig. 4. Uncoded bit error rate (BER) for various 1-bit precoders as a function of the normalized transmit power ϱ and for different antenna configurations and modulation schemes. C1PO and C2PO achieve similar performance to SQUID [6] and significantly outperform linear-quantized precoders, such as quantized zero-forcing (ZF-Q) and MRC (MRC-Q). The performance of ZF precoding with infinite-precision DACs is included as a reference.

For the 16×128 system with 16-QAM, Figure 4(b) shows a similar trend, i.e., non-linear precoders significantly outperform linear-quantized precoders. SQUID outperforms C1PO and C2PO (which perform equally well) by about 0.5 dB SNR at 1% BER. However, we note that the complexity (in terms of operation counts) of SQUID is more than $2\times$ higher than that of C1PO and C2PO, and also involves the sorting of B dimensional vectors which is difficult to implement efficiently in VLSI. We also observe that non-linear precoders enable reliable transmission of higher-order modulation schemes (such as 16-QAM), which is not possible with linear-quantized methods—an observation also made recently in [7].

B. FPGA Implementation Results

To demonstrate the efficacy of C1PO and C2PO, we implemented several FPGA designs for different antenna configurations, namely for 32, 64, 128, and 256 BS antennas; all designs support downlink transmission to 16 UEs for modulation schemes ranging from BPSK to 16-QAM. The FPGA designs were developed on register transfer level (RTL) using Verilog, implemented using Xilinx Vivado Design Suite, and optimized for a Xilinx Virtex-7 XC7VX690T FPGA. Table I shows reference FPGA implementation results for C1PO and C2PO.

We see that the logic area (in terms of slices, logic LUTs, flipflops, and DSP48 units) for all designs increases roughly *linearly* with the number of BS antennas; this is a result of using a linear array of PEs. The only exception is the memory requirements of C1PO (in terms of memory LUTs), which scales roughly *quadratically* in the number of BS antennas; this is a result of having to store the entire $B \times B$ matrix $\mathbf{G}$ in contrast to storing only the augmented $(U+1) \times B$ matrix $\overline{\mathbf{H}}$ for C2PO. We also see that the logic area for C1PO is 20% to 50% smaller than that of C2PO for all array sizes; the memory area of C1PO, however, is significantly larger for 128 and 256 BS antennas. This is because the architecture for C1PO is slightly simpler than that of C2PO, but the memory requirements of C1PO scale quadratically in B whereas the memory requirements of C2PO only scale linearly in B.

The maximum clock frequency for C1PO is slightly higher than that of C2PO, which is due to the slightly simpler architecture of C1PO. As expected, the maximum clock frequency slowly decreases with B, since the FPGA routing overhead increases with B. In fact, after implementing our designs, the critical paths are typically in interconnect networks. Before mapping our designs to the FPGA, however, the critical path for the C1PO designs is in the real-valued multipliers that form part of the complex multiplier, while for the C2PO designs it is in the adders that form part of the complex multiplier. The latency of one C1PO iteration is significantly larger than that of C2PO for 64, 128, and 256 antennas. This results in significantly higher per-iteration throughput of C2PO for these BS antenna array sizes. Hence, C2PO is more efficient (in terms of throughput per area) for large BS antenna arrays, whereas C1PO is more efficient for small ones.

We finally note that the implementation results in Table I ignore the preprocessing complexity. For C1PO, preprocessing requires a $B \times B$ matrix inversion, which is computationally demanding, especially for large BS antenna arrays [20]. In stark contrast, preprocessing for C2PO only requires the computation of the scaled MRC output, which requires a $B \times U$ matrix-vector multiplication. Hence, C2PO can be considered as the preferred method in practical massive MU-MIMO systems.

C. Comparison with MRC-based Precoding

While the papers [26] and [25] propose hardware designs for precoding in massive MU-MIMO systems with high-precision DACs, neither of them provide detailed FPGA implementation results. Reference [26] describes an FPGA-based testbed that uses MRC and ZF-based precoding but does not report area and clock frequency results; [25] only provides ASIC implementation results. To enable a comparison of conventional precoders

TABLE I
IMPLEMENTATION RESULTS FOR C1PO AND C2PO ON A XILINX VIRTEX-7 XC7VX690T FPGA

Algorithm	C1PO				C2PO			
BS antennas B	32	64	128	256	32	64	128	256
Slices	2 700	5 187	10 324	21 951	3 375	6 519	12 690	24 748
LUTs	6 671	13 305	30 979	71 817	10 817	21 920	43 710	85 323
– LUTs as logic	6 031	12 025	25 939	51 897	10 069	20 424	40 718	79 339
– LUTs as memory	640	1 280	5 040	19 920	748	1 496	2 992	5 984
Flipflops	6 830	13 624	26 683	52 175	5 677	12 461	26 083	53 409
DSP48 units	128	256	512	1 024	136	272	544	1 088
Max. clock frequency [MHz]	285	264	244	205	222	206	208	193
Min. latency[a] [clock cycles]	35	67	131	259	39	40	41	42
Max. throughput[a] [Msymbols/s]	130	63	30	13	91	82	81	74
Power consumption[b] [W]	1.13	1.97	3.43	5.74	1.04	1.70	3.17	5.80
Max. throughput/LUTs	19 529	4 733	962	177	8 413	3 756	1 853	862

[a]The latency and maximum throughput are measured for one algorithm iteration.
[b]Statistical power estimation at maximum clock frequency and 1.0 V supply voltage.

TABLE II
IMPLEMENTATION RESULTS FOR A MRC-Q-BASED PRECODER ON A XILINX VIRTEX-7 XC7VX690T FPGA

BS antennas B	32	64	128	256
Slices	2 543	5 097	9 444	17 630
LUTs	7 842	15 617	32 476	64 446
– LUTs as logic	7 010	13 953	29 148	57 790
– LUTs as memory	832	1 664	3 328	6 656
Flipflops	5 711	11 419	21 902	42 764
Clock freq. [MHz]	412	410	388	359
Latency [cycles]	18	18	18	18
TP [Msymbols/s]	366	365	345	319
Power[a] [W]	0.79	1.25	1.84	3.16
Throughput/LUTs	46 665	23 356	10 621	4 945

[a]Statistical power estimation at max. clock freq. and 1.0 V supply voltage.

with C1PO and C2PO, we developed a baseline design that implements MRC followed by quantization (MRC-Q).

Our MRC-Q baseline design is essentially a stripped-down and heavily optimized version of C1PO with only the necessary circuitry to implement MRC-Q. More specifically, our architecture corresponds to B/U linear arrays, each one with U PEs and a control unit. The arrays and PEs are organized as in Figure 2, with the exception that the projection unit is removed from the PEs. In addition, no multipliers are required as MRC-Q computes $\mathbf{H}^H\mathbf{s}$ with $\mathbf{s} \in \mathcal{O}^U$, and hence all multiplications are with constants (given by the constellation set $\mathcal{O}$) and can be implemented with adders and shifters.

The FPGA implementation results for the MRC-Q baseline designs are reported in Table II. Note that these designs do not require any DSP48 units as the multiplication with constants are carried out with conventional logic. In comparison to the 1-bit precoder designs reported in Table I, we see that MRC-Q-based precoding is roughly $5\times$ to $6\times$ more efficient than C2PO and up to $30\times$ more efficient than C1PO (in terms of throughput/LUTs). This efficiency advantage comes at a significant loss in terms of error-rate performance (cf. Figure 4). We note, however, that for massive MU-MIMO systems with significantly more BS antennas than UEs (e.g., more than $8\times$), MRC-Q is a viable low-complexity alternative—a well-known fact in the massive MU-MIMO literature [2]–[4].

D. Performance–Complexity Trade-Offs

In Figure 5, we provide the performance–complexity trade-offs between C1PO (dashed lines with circle markers) and C2PO (dotted lines with square markers) for various BS antenna array sizes. This trade-off is characterized in terms of the minimum normalized transmit power ϱ required to achieve 1% uncoded BER for BPSK (as in Figure 4(a)); the complexity is characterized by the throughput per area (in terms of billion symbols per second per FPGA LUT). As a reference, we also show the performance for ZF precoding with infinite-precision DACs (vertical lines). As in Figure 4, we consider a transmission to $U = 16$ UEs. We see that for small antenna arrays (i.e., for $B = 32$ and $B = 64$), C1PO outperforms C2PO. However, for large antenna arrays (i.e., for $B = 128$ and $B = 256$), C2PO significantly outperforms C1PO. We also see that only a small number of iterations are required (e.g., 2 to 4 iterations) for such large antenna arrays to achieve the performance limits of our algorithms.

In Figure 5, we additionally show the trade-off achieved by the MRC-Q baseline design reported in Section VI-C. Clearly, MRC-Q achieves higher throughput per LUT than C1PO and C2PO for large BS arrays ($B = 128$ and $B = 256$); this gain comes, however, at the cost of rather poor error-rate

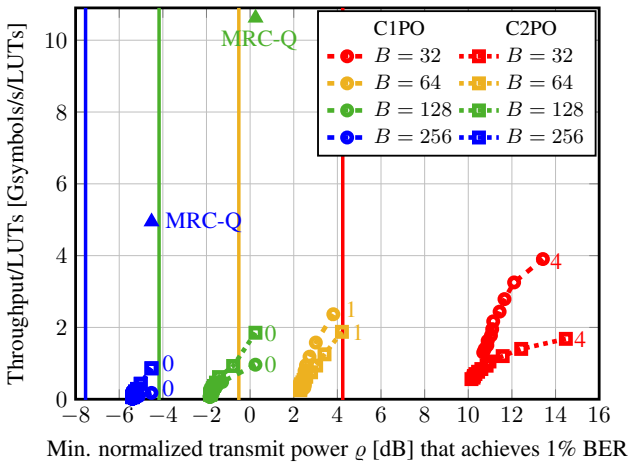


Fig. 5. Performance–complexity trade-offs for C1PO and C2PO. The numbers next to the curves correspond to the number of iterations $t_{\max}$. For $t_{\max} = 0$, we directly take the outputs from the initialization step $\mathbf{x}^{(1)} = \mathbf{H}^H\mathbf{s}$, which is an approach equivalent to MRC-Q. The vertical lines show the performance of ZF precoding with infinite-precision DACs. C1PO outperforms C2PO for small BS antenna arrays ($B = 32$ and $B = 64$); C2PO outperforms C1PO for large antenna arrays ($B = 128$ and $B = 256$). MRC-Q achieves higher throughput per LUT at the cost of rather poor performance.

performance. For small BS antenna arrays ($B = 32$ and $B = 64$), MRC-Q is unable to achieve the target BER of 1%. Hence, MRC-Q is only suitable for massive MU-MIMO systems with very high BS-to-UE-antenna ratios in which best-in-class error-rate performance is not the main design objective.

VII. Conclusions

We have proposed two nonlinear precoding algorithms, namely C1PO and C2PO, which achieve excellent error-rate performance in 1-bit massive MU-MIMO systems at low computational complexity. To substantiate this claim, we have designed corresponding VLSI architectures—to the best of our knowledge, the first for 1-bit massive MU-MIMO systems—and have presented FPGA reference implementations for a variety of BS antenna array configurations. Our results demonstrate that nonlinear precoding for 1-bit massive MU-MIMO systems is feasible from a hardware implementation perspective, even for antenna arrays with hundreds of BS antennas. As a result, our hardware designs enable BS antenna arrays with 1-bit DACs to reliably transmit high-rate data to multiple UEs, which has the potential to keep the the hardware complexity, system costs, and circuit power consumption within manageable limits.

There are many avenues for future work. Besides our convergence results, a theoretical performance analysis of C1PO and C2PO is a challenging open research topic. Implementing precoders for other nonlinear algorithms, such as SQUID [6], which perform better than C1PO and C2PO at low SNR, is left for future work. The design of 1-bit precoding algorithms and hardware accelerators for wideband massive MU-MIMO systems that use orthogonal-frequency division multiplexing (OFDM) is the subject of ongoing work.

Appendix A
Proof of Theorem 1

Let $E(\mathbf{z},\mathbf{x}) = \|\mathbf{A}\mathbf{z}\|_2^2 + \gamma\|\mathbf{z}-\mathbf{x}\|_2^2 - \delta\|\mathbf{x}\|_2^2$ denote the objective (7) minimized by C1PO. Because $\mathcal{B}^B$ is bounded, the sequence of iterates $\{(\mathbf{z}^{(t)},\mathbf{x}^{(t)})\}$ remains bounded and thus contains a convergent sub-sequence. Denote the limit of this sub-sequence by $(\mathbf{z}^\star,\mathbf{x}^\star)$ and set $E^\star = E(\mathbf{z}^\star,\mathbf{x}^\star)$. Consider the point $\hat{\mathbf{z}}^\star = \arg\min_{\mathbf{z}} E(\mathbf{z},\mathbf{x}^\star) = (\mathbf{I}_B + \gamma^{-1}\mathbf{A}^H\mathbf{A})^{-1}\mathbf{x}^\star$. If $\hat{\mathbf{z}}^\star \neq \mathbf{z}^\star$, then we have the strict inequality

$$E((\hat{\mathbf{z}}^\star + \mathbf{z}^\star)/2,\mathbf{x}^\star) < \frac{1}{2}E(\hat{\mathbf{z}}^\star,\mathbf{x}^\star) + \frac{1}{2}E(\mathbf{z}^\star,\mathbf{x}^\star) = E^\star$$

because E is strongly convex in $\mathbf{z}$. However, this contradicts the fact that $\hat{\mathbf{z}}^\star = \arg\min_{\mathbf{z}} E(\mathbf{z},\mathbf{x}^\star)$, and so it must be the case that $\hat{\mathbf{z}}^\star = \mathbf{z}^\star$. Because $\delta < \gamma$, E is strongly convex in $\mathbf{x}$, and a similar argument shows that $\mathbf{x}^\star = \arg\min_{\mathbf{x}\in\mathcal{B}^B} E(\mathbf{z}^\star,\mathbf{x})$. Hence, $(\mathbf{z}^\star,\mathbf{x}^\star)$ minimizes E with respect to $\mathbf{z}$ and $\mathbf{x}$ separately; this, combined with the fact that E is differentiable, and $\mathcal{B}$ coordinate-wise separable, guarantees that $(\mathbf{z}^\star,\mathbf{x}^\star)$ satisfies the first-order conditions for (7); see Theorem 2 in [39] and similar arguments in [40].

Appendix B
Proof of Theorem 2

Let $E(\mathbf{z},\mathbf{x}) = \|\mathbf{A}\mathbf{x}\|_2^2 - \delta\|\mathbf{x}\|_2^2$ denote the objective (11) minimized by C2PO. Let f and g be defined as in (12). Using the definition of the proximal operator (13) together with (15), the second update (14) of C2PO can be written as

$$\begin{aligned}\mathbf{x}^{(t+1)} &= \arg\min_{\mathbf{x}} g(\mathbf{x}) + \frac{1}{2\tau}\|\mathbf{x} - (\mathbf{x}^{(t)} - \tau\nabla f(\mathbf{x}^{(t)}))\|^2 \\ &= \arg\min_{\mathbf{x}} g(\mathbf{x}) + f(\mathbf{x}^{(t)}) + \langle \mathbf{x}-\mathbf{x}^{(t)}, \nabla f(\mathbf{x}^{(t)})\rangle \\ &\qquad + \frac{1}{2\tau}\|\mathbf{x}-\mathbf{x}^{(t)}\|^2.\end{aligned}$$

Observe that, whenever $\tau < \|\mathbf{A}^T\mathbf{A}\|_{2,2}^{-1}$, the inequality

$$f(\mathbf{x}) \leq f(\mathbf{x}^{(t)}) + \langle \mathbf{x}-\mathbf{x}^{(t)}, \nabla f(\mathbf{x}^{(t)})\rangle + \frac{1}{2\tau}\|\mathbf{x}-\mathbf{x}^{(t)}\|^2$$

holds for all $\mathbf{x}$. Using this observation, we can write

$$\begin{aligned}E(\mathbf{x}^{(t+1)}) &= g(\mathbf{x}^{(t+1)}) + f(\mathbf{x}^{(t+1)}) \\ &\leq g(\mathbf{x}^{(t+1)}) + f(\mathbf{x}^{(t)}) + \langle \mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}, \nabla f(\mathbf{x}^{(t)})\rangle \\ &\quad + \frac{1}{2\tau}\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\|^2 \\ &= \min_{\mathbf{x}} g(\mathbf{x}) + f(\mathbf{x}^{(t)}) + \langle \mathbf{x}-\mathbf{x}^{(t)}, \nabla f(\mathbf{x}^{(t)})\rangle \\ &\qquad + \frac{1}{2\tau}\|\mathbf{x}-\mathbf{x}^{(t)}\|^2 \\ &\leq g(\mathbf{x}^{(t)}) + f(\mathbf{x}^{(t)}) = E(\mathbf{x}^{(t)}).\end{aligned}$$

This shows that the sequence $\{E(\mathbf{x}^{(t)})\}$ is monotonically decreasing. Since the sequence is bounded below, there is some limit $L = \lim_{t\to\infty} E(\mathbf{x}^{(t)})$. Let $\{\mathbf{x}^{(t_k)}\}$ be a convergent sub-sequence of iterates (which must exist because the iterates are bounded) with limit point $\mathbf{x}^\star$. Let

$$\begin{aligned}\bar{\mathbf{x}}^\star = \arg\min_{\mathbf{x}}\ & g(\mathbf{x}) + f(\mathbf{x}^\star) + \langle \mathbf{x}-\mathbf{x}^\star, \nabla f(\mathbf{x}^\star)\rangle \\ & + \frac{1}{2\tau}\|\mathbf{x}-\mathbf{x}^\star\|^2 \end{aligned} \tag{17}$$

be the result of applying the C2PO iteration starting at $\mathbf{x}^\star$. Observing that $E(\mathbf{x}^{(t_k+1)}) \leq E(\mathbf{x}^{(t_k)}) \leq E(\mathbf{x}^{(t_k-1)})$, and letting $k \to \infty$, we find that $E(\bar{\mathbf{x}}^\star) = E(\mathbf{x}^\star) = L$, and

so $\mathbf{x}^\star$ is a minimizer of (17). This is only possible if $0 \in \partial g(\mathbf{x}^\star) + \nabla f(\mathbf{x}^\star)$, in which case $\mathbf{x}^\star$ is a stationary point.

ACKNOWLEDGMENTS

The authors thank O. Tirkkonen for insightful discussions on 1-bit precoding. The work of O. Castañeda and C. Studer was supported in part by Xilinx Inc. and by the US National Science Foundation (NSF) under grants ECCS-1408006 and CCF-1535897. The work of S. Jacobsson and G. Durisi was supported by the Swedish Foundation for Strategic Research under grant ID14-0022, and by the Swedish Governmental Agency for Innovation Systems (VINNOVA) within the center ChaseOn. The work of T. Goldstein was supported in part by the US NSF under grant CCF-1535902 and by the US Office of Naval Research under grant N00014-17-1-2078.

REFERENCES

[1] O. Castañeda, T. Goldstein, and C. Studer, "POKEMON: a non-linear beamforming algorithm for 1-bit massive MIMO," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, New Orleans, LA, Mar. 2017.

[2] F. Rusek, D. Persson, B. Kiong, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Oppurtunities and challenges with very large large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[3] E. G. Larsson, F. Tufvesson, O. Edfors, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[4] L. Lu, G. Ye Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.

[5] K. Li, R. Sharan, Y. Chen, T. Goldstein, J. R. Cavallaro, and C. Studer, "Decentralized beamforming for massive MU-MIMO on a GPU cluster," in *4th IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Washington, D.C., Dec. 2016.

[6] S. Jacobsson, G. Durisi, M. Coldrey, T. Goldstein, and C. Studer, "Quantized precoding for massive MU-MIMO," *arXiv preprint: 1610.07564*, Oct. 2016.

[7] ——, "Nonlinear 1-bit precoding for massive MU-MIMO with higher-order modulation," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, Nov. 2016.

[8] H. Jedda, J. A. Nossek, and A. Mezghani, "Minimum BER precoding in 1-bit massive MIMO systems," in *IEEE Sensor Array and Multichannel Signal Process. Workshop (SAM)*, Rio de Janeiro, Brazil, Jul. 2016.

[9] O. Tirkkonen and C. Studer, "Subset-codebook precoding for 1-bit massive multiuser MIMO," in *Conference on Information Sciences and Systems*, Baltimore, Maryland, Mar. 2017.

[10] A. Mezghani, R. Ghiat, and J. A. Nossek, "Transmit processing with low resolution D/A-converters," in *Proc. IEEE Int. Conf. Electron., Circuits, Syst. (ICECS)*, Yasmine Hammamet, Tunisia, Dec. 2009, pp. 683–686.

[11] A. K. Saxena, I. Fijalkow, and A. L. Swindlehurst, "On one-bit quantized ZF precoding for the multiuser massive MIMO downlink," in *IEEE Sensor Array and Multichannel Signal Process. Workshop (SAM)*, Rio de Janeiro, Brazil, Jul. 2016.

[12] R. D. J. Guerreiro and P. Montezuma, "Use of 1-bit digital-to-analogue converters in massive MIMO systems," *IEEE Electron. Lett.*, vol. 52, no. 9, pp. 778–779, Apr. 2016.

[13] O. B. Usman, H. Jedda, A. Mezghani, and J. A. Nossek, "MMSE precoder for massive MIMO using 1-bit quantization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 3381–3385.

[14] S. Shah, A. K. Yadav, C. D. Castillo, D. W. Jacobs, C. Studer, and T. Goldstein, "Biconvex relaxation for semidefinite programming in computer vision," in *European Conference on Computer Vision (ECCV)*. Springer, Sep. 2016, pp. 717–735.

[15] C. Risi, D. Persson, and E. G. Larsson, "Massive MIMO with 1-bit ADC," Apr. 2014. [Online]. Available: http://arxiv.org/abs/1404.7736

[16] S. Jacobsson, G. Durisi, M. Coldrey, U. Gustavsson, and C. Studer, "One-bit massive MIMO: Channel estimation and high-order modulations," in *Proc. IEEE Int. Conf. Commun. Workshop (ICCW)*, London, U.K., June 2015, pp. 1304–1309.

[17] Y. Li, C. Tao, G. Seco-Granados, A. Mezghani, A. L. Swindlehurst, and L. Liu, "Channel estimation and performance analysis of one-bit massive MIMO systems," Sep. 2016. [Online]. Available: https://arxiv.org/abs/1609.07427

[18] C. Mollén, J. Choi, E. G. Larsson, and R. W. Heath Jr., "Performance of the wideband massive uplink MIMO with one-bit ADCs," Feb. 2016. [Online]. Available: https://arxiv.org/abs/1602.07364

[19] C. Studer and G. Durisi, "Quantized massive MU-MIMO-OFDM uplink," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2387–2399, Jun. 2016.

[20] M. Wu, B. Yin, G. Wang, C. Dick, J. Cavallaro, and C. Studer, "Large-scale MIMO detection for 3GPP LTE: Algorithm and FPGA implementation," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 916–929, Oct. 2014.

[21] Z. Wu, C. Zhang, Y. Xue, S. Xu, and X. You, "Efficient architecture for soft-output massive MIMO detection with Gauss-Seidel method," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, Montreal, Canada, Aug. 2016, pp. 1886–1889.

[22] M. Wu, C. Dick, J. R. Cavallaro, and C. Studer, "High-throughput data detection for massive MU-MIMO-OFDM using coordinate descent," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 63, no. 12, pp. 2357–2367, Nov. 2016.

[23] O. Castañeda, T. Goldstein, and C. Studer, "Data detection in large multi-antenna wireless systems via approximate semidefinite relaxation," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 63, no. 12, pp. 2334–2346, Nov. 2016.

[24] M. Barrenechea, L. Barbero, M. Mendicute, and J. Thompson, "Design and hardware implementation of a low-complexity multiuser vector precoder," in *Conference on Design and Architectures for Signal and Image Processing (DASIP)*, Oct. 2010, pp. 160–167.

[25] H. Prabhu, O. Edfors, J. Rodrigues, L. Liu, and F. Rusek, "Hardware efficient approximative matrix inversion for linear pre-coding in massive MIMO," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, June 2014, pp. 1700–1703.

[26] C. Shepard, N. Anand, and L. Zhong, "Practical performance of MU-MIMO precoding in many-antenna base stations," in *Proc. of the 2013 workshop on Cellular networks: operations, challenges, and future design*. ACM, June 2013, pp. 13–18.

[27] E. Björnson, M. Bengtsson, and B. Ottersten, "Optimal multiuser transmit beamforming: A difficult problem with a simple solution structure," *IEEE Signal Process. Mag.*, vol. 31, no. 4, pp. 142–148, Jul. 2014.

[28] E. Björnson and E. Jorswieck, "Optimal resource allocation in coordinated multi-cell systems," *Foundations and Trends in Communications and Information Theory*, vol. 9, no. 2-3, pp. 113–381, 2013.

[29] M. Joham, W. Utschick, and J. A. Nossek, "Linear transmit processing in MIMO communications systems," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2700–2712, Aug. 2005.

[30] S. Shi, M. Schubert, and H. Boche, "Downlink MMSE transceiver optimization for multiuser MIMO systems: Duality and sum-MSE minimization," *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5436–5446, Nov. 2007.

[31] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger, "Closest point search in lattices," *IEEE Trans. Inf. Theory*, vol. 48, no. 8, pp. 2201–2214, Aug. 2002.

[32] U. Fincke and M. Pohst, "Improved methods for calculating vectors of short length in a lattice, including a complexity analysis," *Math. Comput.*, vol. 44, no. 170, pp. 463–471, Apr. 1985.

[33] S. Verdú, "Computational complexity of multiuser detection," *Algorithmica*, vol. 4, no. 1, pp. 303–312, 1989.

[34] T. Goldstein, C. Studer, and R. G. Baraniuk, "A field guide to forward-backward splitting with a FASTA implementation," Nov. 2014. [Online]. Available: http://arxiv.org/abs/1411.3406

[35] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, Jan. 2009.

[36] T. Goldstein and S. Setzer, "High-order methods for basis pursuit," *UCLA CAM Report*, pp. 10–41, 2010.

[37] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends® in Optimization*, vol. 1, no. 3, pp. 127–239, Jan. 2014.

[38] G. H. Golub and C. F. van Loan, *Matrix Computations*, 3rd ed. The Johns Hopkins Univ. Press, 1996.

[39] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *Journal of optimization theory and applications*, vol. 109, no. 3, pp. 475–494, 2001.

[40] P. Richtárik and M. Takáč, "Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function," *Mathematical Programming*, vol. 144, no. 1-2, pp. 1–38, 2014.