VLSI Design of a Nonparametric Equalizer for Massive MU-MIMO

Charles Jeon[1], Gulnar Mirza[1], Ramina Ghods[1], Arian Maleki[2], and Christoph Studer[1]

[1]Cornell University, Ithaca, NY; {cj339, gzm3, rg548, studer}@cornell.edu; web: http://vip.ece.cornell.edu
[2]Columbia University, New York City, NY; arian@stat.columbia.edu

Abstract—Linear minimum mean-square error (L-MMSE) equalization is among the most popular methods for data detection in massive multi-user multiple-input multiple-output (MU-MIMO) wireless systems. While L-MMSE equalization enables near-optimal spectral efficiency, accurate knowledge of the signal and noise powers is necessary. Furthermore, corresponding VLSI designs must solve linear systems of equations, which requires high arithmetic precision, exhibits stringent data dependencies, and results in high circuit complexity. This paper proposes the first VLSI design of the NOnParametric Equalizer (NOPE), which avoids knowledge of the transmit signal and noise powers, provably delivers the performance of L-MMSE equalization for massive MU-MIMO systems, and is resilient to numerous system and hardware impairments due to its parameter-free nature. Moreover, NOPE avoids computation of a matrix inverse and only requires hardware-friendly matrix-vector multiplications. To showcase the practical advantages of NOPE, we propose a parallel VLSI architecture and provide synthesis results in 28 nm CMOS. We demonstrate that NOPE performs on par with existing data detectors for massive MU-MIMO that require accurate knowledge of the signal and noise powers.

I. INTRODUCTION

It is widely believed that massive multi-user multiple-input multiple-output (MU-MIMO) will be a core technology for fifth-generation (5G) wireless systems. Massive MU-MIMO relies on base-station (BS) architectures with hundreds of antenna elements and radio-frequency (RF) chains that serve tens of user equipments (UEs) in the same time-frequency resource. While this emerging technology enables unprecedented spectral efficiency by means of fine-grained beamforming [1], [2], it also poses significant practical implementation challenges.

A. The Case for Nonparametric Equalization

Data detection in the uplink (UEs transmit data to the BS) is among the most critical tasks from a spectral efficiency and hardware complexity perspective [3]. While optimal MIMO data detection is known to be NP-hard [4], it has been shown in [3], [5] that linear minimum mean-square error (L-MMSE) equalization enables near-optimal performance in massive MU-MIMO systems. However, L-MMSE equalization requires accurate knowledge of the signal and noise powers [6]. Furthermore, corresponding hardware designs must solve linear systems

of equations, which requires high arithmetic precision and suffers from stringent data dependencies—both of these aspects result in relatively high circuit complexity [7]–[11]. In addition, practical massive MU-MIMO BS designs will most likely rely on inexpensive RF circuitry which suffers from numerous impairments, including amplifier nonlinearities, phase noise, and quantization artifacts [12], [13]. The presence of non-ideal hardware necessitates the design of new equalization algorithms that are resilient to real-world hardware imperfections.

Recently, a novel algorithm called *NonParametric Equalizer* (NOPE, for short) was proposed in [6]. NOPE does not require knowledge of the signal and noise powers while provably achieving the performance of L-MMSE equalization in massive MU-MIMO systems. NOPE combines approximate message passing (AMP) [14] with Stein's unbiased risk estimator (SURE) [15] and mismatched data detection [16], which renders this algorithm resilient to numerous hardware impairments while being computationally efficient: NOPE only requires matrix-vector products and avoids a computation of costly matrix inverses or matrix decompositions, which are typically required by L-MMSE equalizer algorithms. Despite all these advantages, NOPE has been designed only for idealistic channel models and has not yet been integrated in hardware.

B. Contributions

In this paper, we generalize NOPE to practical channels and provide, to the best of our knowledge, its first VLSI design. Our contributions are summarized as follows:

- We propose a set of algorithm-level modifications that enable NOPE to operate on more realistic MU channels.
- We develop a VLSI architecture that relies on Cannon's algorithm [17] to achieve high throughput at low area.
- We show reference VLSI synthesis results in 28 nm CMOS for a 64 BS antenna, 16 UE massive MU-MIMO system.
- We compare NOPE to existing massive MU-MIMO equalizers requiring knowledge of the signal and noise powers.

Our results demonstrate that massive MU-MIMO has the unique potential to design parameter-free algorithms, such as NOPE, that perform on par with solutions that require accurate knowledge of critical system and model parameters.

C. Notation

Lowercase and uppercase boldface letters designate column vectors and matrices, respectively. The transpose and conjugate

transpose of a matrix $\mathbf{A}$ are denoted by $\mathbf{A}^\mathrm{T}$ and $\mathbf{A}^\mathrm{H}$, respectively; the entry on the kth row and ℓth column of $\mathbf{A}$ is $A_{k,\ell}$. The $M \times M$ identity matrix is denoted by $\mathbf{I}_M$ and the $M \times N$ all-zeros matrix by $\mathbf{0}_{M \times N}$. The kth entry of a vector $\mathbf{a}$ is a_k. We define the Hadamard product as $\circ$. For an N-dimensional vector $\mathbf{a}$, we define $\langle \mathbf{a} \rangle = \frac{1}{N}\sum_{k=1}^{N} a_k$. The probability density function (PDF) of a circularly-symmetric complex-valued Gaussian random vector with covariance matrix $\mathbf{K}$ is denoted by $\mathcal{CN}(\mathbf{0}, \mathbf{K})$. Expectation and variance with respect to the random vector $\mathbf{a}$ is denoted by $\mathbb{E}_\mathbf{a}[\cdot]$ and $\mathbb{V}\mathrm{ar}_\mathbf{a}[\cdot]$, respectively.

II. A PRIMER ON L-MMSE EQUALIZATION

We now introduce the system model and review the basics of L-MMSE equalization. We then discuss NOPE.

A. System Model

We consider the input-output relation $\mathbf{y} = \mathbf{Hx} + \mathbf{n}$ to model a massive MU-MIMO uplink system operating in a frequency-flat channel [3]. The vector $\mathbf{y} \in \mathbb{C}^B$ contains the received signals at the BS; B denotes the number of BS antennas; the matrix $\mathbf{H} \in \mathbb{C}^{B \times U}$ represents the uplink MIMO channel; U denotes the number of UEs; the transmit signal vector is $\mathbf{x} \in \mathbb{C}^U$; and the vector $\mathbf{n} \in \mathbb{C}^B$ models receive noise, which has i.i.d. circularly-symmetric complex Gaussian entries with variance N_0 per entry. Throughout this paper, we assume that the transmit signal vector $\mathbf{x}$ has i.i.d. entries so that $p(\mathbf{x}) = \prod_{u=1}^{U} p(x_u)$, where $p(\cdot)$ models the signal prior (e.g., a 16-QAM constellation) with zero mean and signal variance $E_\mathrm{x} = \mathbb{E}\left[|x_u|^2\right]$, $u = 1, \ldots, U$. We require the following definitions.

Definition 1. *The* antenna ratio *is defined as $\beta = U/B$.*

Definition 2. *The* large-antenna limit *is defined by fixing the antenna ratio β and letting $U \to \infty$.*

Definition 3. *The channel matrix $\mathbf{H}$ is said to have* uniform channel gains *if the entries are i.i.d. circularly-symmetric complex Gaussian with variance $1/B$ per complex entry.*

B. Basics of L-MMSE Equalization

L-MMSE equalization is among the most popular methods to compute an estimate for $\mathbf{x}$ from $\mathbf{y}$ and from knowledge of the channel matrix $\mathbf{H}$, and enjoys widespread use for data detection in MIMO systems [7], [18]–[20]. The relatively low computational complexity (except for the inversion of a potentially large matrix) and acceptable performance render this method a feasible alternative to more complicated data detection algorithms. Moreover, it has been shown in [3], [5] that L-MMSE equalization enables (often significantly) higher achievable rates than zero-forcing (ZF) or maximum ratio combining (MRC)-based equalizers in massive MU-MIMO systems. However, to enable near-optimal spectral efficiency via L-MMSE equalization, accurate knowledge of the signal and noise powers is required; see, e.g., [6], [21].

Mathematically, the goal of L-MMSE equalization is to compute a linear estimate $\hat{\mathbf{x}} = \mathbf{Wy}$ from the receive vector $\mathbf{y}$ that minimizes the $MSE = \mathbb{E}_{\mathbf{x},\mathbf{n}}\left[\|\hat{\mathbf{x}} - \mathbf{x}\|^2\right]$ using knowledge of the channel matrix $\mathbf{H}$ as well as the signal and noise powers,

E_x and N_0, respectively. For a circularly-symmetric complex-valued transmit signal $\mathbf{x}$, the equalization matrix $\mathbf{W} \in \mathbb{C}^{U \times B}$ is given by $\mathbf{W} = (\mathbf{H}^\mathrm{H}\mathbf{H} + \frac{N_0}{E_\mathrm{x}}\mathbf{I}_U)^{-1}\mathbf{H}^\mathrm{H}$. If the signal $\mathbf{x}$ is zero-mean and real-valued (e.g., for BPSK signals), then the optimal linear estimator for the real part $\mathbf{x}_\mathrm{Re}$ is given by

$$\hat{\mathbf{x}}_\mathrm{Re} = \left(\mathbf{H}_\mathrm{Re}^\mathrm{T}\mathbf{H}_\mathrm{Re} + \mathbf{H}_\mathrm{Im}^\mathrm{T}\mathbf{H}_\mathrm{Im} + \frac{N_0}{2E_\mathrm{x}}\mathbf{I}_U\right)^{-1}(\mathbf{H}_\mathrm{Re}^\mathrm{T}\mathbf{y}_\mathrm{Re} + \mathbf{H}_\mathrm{Im}^\mathrm{T}\mathbf{y}_\mathrm{Im}),$$

where $\mathbf{H}_\mathrm{Re}$, $\mathbf{H}_\mathrm{Im}$, $\mathbf{y}_\mathrm{Re}$, and $\mathbf{y}_\mathrm{Im}$ are the real and imaginary parts of $\mathbf{H}$ and $\mathbf{y}$, respectively; the imaginary part of the estimate is $\hat{\mathbf{x}}_\mathrm{Im} = \mathbf{0}_{U \times 1}$. Clearly, L-MMSE equalization relies on knowledge of the quantities $\rho = N_0/E_\mathrm{x}$ or $\rho = N_0/(2E_\mathrm{x})$, which requires (i) means to detect whether the transmit signals are real- or complex-valued and (ii) an accurate estimate of ρ that is commonly acquired in a dedicated training phase [22].

C. L-MMSE Equalization via AMP

As shown in [16], L-MMSE equalization can be implemented using the mismatched complex Bayesian AMP (mcB-AMP) framework. By assuming a mismatched Gaussian signal prior distribution $\tilde{p}(\mathbf{x}) = \prod_{i=u}^{U} \tilde{p}(x_u)$ with $\tilde{p}(x_u) \sim \mathcal{CN}(0, E_\mathrm{x})$ instead of the true signal prior $X_0 \sim p(x_0)$ (e.g., two Dirac delta functions concentrated at -1 and $+1$ for BPSK), one can design the following *parametric* L-MMSE algorithm:

Algorithm 1 (L-MMSE-AMP [16]). *Initialize $t = 1$, $x_u^1 = 0$, $u = 1, \ldots, U$, and $\mathbf{r}^1 = \mathbf{y} - \mathbf{Hx}^1$. Then, for every iteration $t = 1, \ldots, t_\mathrm{max}$ compute the output $\mathbf{z}^t$ via the following steps:*

$$\tilde{\sigma}_t^2 = \frac{1}{B}\|\mathbf{r}^t\|_2^2$$
$$\tau^t = \arg\min_{\tau \geq 0} \Psi(\tilde{\sigma}_t^2, \tau) \qquad (1)$$
$$\mathbf{z}^t = \mathbf{x}^t + \mathbf{H}^\mathrm{H}\mathbf{r}^t$$
$$\mathbf{x}^{t+1} = \mathsf{F}^{mm}(\mathbf{z}^t, \tau^t) \qquad (2)$$
$$\mathbf{r}^{t+1} = \mathbf{y} - \mathbf{Hx}^{t+1} + \beta\mathbf{r}^t\langle\mathsf{F}'^{mm}(\mathbf{z}^t, \tau^t)\rangle. \qquad (3)$$

Here, the posterior mean function $\mathsf{F}^{mm}(x_u, \tau) = \frac{E_\mathrm{x}}{E_\mathrm{x}+\tau}x_u$ and $\mathsf{F}'^{mm}(x_u, \tau) = \frac{E_\mathrm{x}}{E_\mathrm{x}+\tau}$ operate element-wise on vectors. We furthermore need the MSE function: $\Psi(\tilde{\sigma}_t^2, \tau) = \frac{\tau^2 E_\mathrm{x} + \tilde{\sigma}_t^2 E_\mathrm{x}^2}{(E_\mathrm{x}+\tau)^2}$.

Interestingly, the estimate $\hat{\mathbf{z}} = \lim_{t\to\infty} \mathbf{z}^t$ computed by L-MMSE-AMP exhibits the same MSE as that of the L-MMSE equalizer in the large-antenna limit and for matrices $\mathbf{H}$ with uniform channel gains [16]. While this is an asymptotic equivalence, reference [6] has shown that the error-rate performance of L-MMSE-AMP is virtually indistinguishable from an L-MMSE equalizer in practical (finite-dimensional) massive MU-MIMO systems for a small number of iterations t_max (ten or fewer). Clearly, Algorithm 1 mainly relies on matrix-vector multiplications, which enables parallel hardware designs. However, the exact knowledge of E_x is still required.

III. NONPARAMETRIC EQUALIZER: NOPE

We now summarize the necessary steps to free Algorithm 1 from knowledge of the signal power, leading to NOPE. We then propose a generalization of the algorithm that makes it suitable for more realistic MIMO system scenarios.

A. The NOPE Algorithm

To develop NOPE, we wish to automatically tune the signal power E_x and the parameter τ^t. To this end, we first introduce the parameter $\gamma^t = E_\mathrm{x}/\tau^t$ and reparametrize the functions $\mathsf{F}^{\mathrm{mm}}(x_u, \gamma^t) = \frac{\gamma^t}{\gamma^t+1}x_u$ and $\mathsf{F}'^{\mathrm{mm}}(x_u, \gamma^t) = \frac{\gamma^t}{\gamma^t+1}$ in Algorithm 1. Now, only a single parameter must be tuned per iteration, i.e., γ^t. Interestingly, [16, Thm. 3] shows that optimal parameter tuning is achieved by tuning each parameter γ^t by minimizing (1) separately at iteration t starting from $t = 1$ to t_max. Hence, the remaining piece is to replace the MSE function Ψ with a function that does not depend on the true signal prior $p(x_0)$. As shown in [6], one can use Stein's unbiased risk estimate (SURE) [15] to extract an estimate of the MSE function Ψ as

$$\hat{\Psi}(\tilde{\sigma}_t^2, \gamma^t) = \tilde{\sigma}_t^2 \frac{\gamma^t-1}{\gamma^t+1} + \frac{\|\mathbf{z}^t\|_2^2}{U(\gamma^t+1)^2}. \qquad (4)$$

Since the minimum of $\hat{\Psi}$ is given by $\gamma_\mathrm{min}^t = \|\mathbf{z}^t\|_2^2/(U\tilde{\sigma}_t^2) - 1$ we can replace the tuning stage in (1) by γ_min^t, which leads to NOPE. As proven in [6, Cor. 6], NOPE achieves the performance of an L-MMSE equalizer in the large antenna limit given that $\mathbf{H}$ has uniform channel gains and for $t \to \infty$.

B. Robust Version of NOPE

NOPE and Algorithm 1 require the matrix $\mathbf{H}$ to have uniform channel gains. However, in practice each UE typically has a different large-scale fading gain (e.g., affected by the distance to the BS), resulting in channel matrices $\mathbf{H}$ whose columns have different scale. We now show how NOPE can be made robust to such channels. As in [6], one can rewrite the channel matrix as $\mathbf{H} = \tilde{\mathbf{H}}\mathbf{D}$, where each element of $\tilde{\mathbf{H}}$ is distributed as $\mathcal{CN}(0, 1/B)$ and $\mathbf{D}$ is diagonal containing the uth UE's individual large-scale fading gain d_u. For this model, one must estimate the gain of the uth UE by $\sum_{b=1}^{B}|H_{b,u}|^2 = d_u^2 \sum_{b=1}^{B}|\tilde{H}_{b,u}|^2$, which converges to d_u^2 in the large-antenna limit. Thus, $\mathbf{D}$ is estimated with a diagonal matrix $\hat{\mathbf{D}}$, where the uth diagonal element is given by $\hat{d}_u$. To enable NOPE to support nonuniform channel gains, we modify the posterior mean function in (2) into an element-wise operation [6]

$$\mathsf{F}_u^{\mathrm{mm}}(z_u^t, \tau^t) = \frac{E_\mathrm{x}}{E_\mathrm{x} + \tau^t/\hat{d}_u^2} z_u^t. \qquad (5)$$

Furthermore, step (3) in Algorithm 1 must be replaced by

$$\mathbf{r}^{t+1} = \mathbf{y} - \mathbf{H}\mathbf{x}^{t+1} + \beta\mathbf{r}^t \frac{1}{U}\sum_{u=1}^{U}\mathsf{F}_u'^{\mathrm{mm}}(z_u^t, \tau^t)$$

to take into account the fact that different functions $\mathsf{F}_u^{\mathrm{mm}}(z_u^t, \tau^t)$ are used for each UE. This generalization also requires new estimates for the parameters E_x and τ^t in (5). As shown in [6, Thm 7], both of these parameters can be estimated as follows

$$\hat{E}_\mathrm{x}^t = \frac{v_\mathbf{z}^t - 2v_\mathbf{r}^t}{\sum_{u=1}^{U}\hat{d}_u^2} \quad \text{and} \quad \hat{\tau}^t = \frac{1}{B}\|\mathbf{r}^t\|^2, \qquad (6)$$

where we introduced shorthand notation $v_\mathbf{z}^t = \sum_{u=1}^{U}\hat{d}_u^2|z_u^t|^2$ for the weighted-norm of $\mathbf{z}^t$ with respect to its large-scale fading gains, and $v_\mathbf{r}^t = \beta\|\mathbf{r}\|_2^2/2$ for the residual norm.

The remaining piece of our robust NOPE is to enable L-MMSE data detection for BPSK constellations for which

---

Algorithm 2 Robust version of NOPE
1: **inputs:** $\mathbf{H} \in \mathbb{C}^{B \times U}$ and $\mathbf{y} \in \mathbb{C}^B$
2: **precompute:** $\hat{d}_u^2 = \sum_{b=1}^{B}|H_{b,u}|^2$, $\hat{d}_u^{-2} = 1/\hat{d}_u^2$ for all $u = 1, \ldots, U$, and $\langle\hat{\mathbf{d}}^2\rangle = \frac{1}{U}\sum_{u=1}^{U}\hat{d}_u^2$
3: **initialize:** $t = 1$, $\mathbf{x}^1 = \mathbf{0}_{U \times 1}$, and $\langle\boldsymbol{\alpha}\rangle = 0$
4: **for** $t = 1, 2, \ldots, t_\mathrm{max}$ **do**
5:    $\mathbf{r}^t = \mathbf{y} - \mathbf{H}\mathbf{x}^t + \frac{\beta}{2}\langle\boldsymbol{\alpha}\rangle\mathbf{r}^t$ (residual update)
6:    $v_\mathbf{r}^t = \frac{\beta}{2}\|\mathbf{r}^t\|_2^2$ (norm of $\mathbf{r}^t$)
7:    $\mathbf{z}^t = \mathbf{x}^t + \hat{\mathbf{d}}^{-2} \circ (\mathbf{H}^H\mathbf{r}^t)$
8:    $v_{\mathbf{z},\mathrm{Re}}^t = \sum_{u=1}^{U}\hat{d}_u^2\mathrm{Re}\{z_u^t\}^2$ (weighted norm of $\mathrm{Re}\{\mathbf{z}^t\}$)
9:    $v_{\mathbf{z},\mathrm{Im}}^t = \sum_{u=1}^{U}\hat{d}_u^2\mathrm{Im}\{z_u^t\}^2$ (weighted norm of $\mathrm{Im}\{\mathbf{z}^t\}$)
10:   $K^t = (v_\mathbf{r}^t\langle\hat{\mathbf{d}}^2\rangle)^{-1}$
11:   **for** $u = 1, \ldots, U$ **do**
12:      $\alpha_{u,\mathrm{Re}}^t = (1 + (K^t\hat{d}_u^2(v_{\mathbf{z},\mathrm{Re}}^t - v_\mathbf{r}^t))^{-1})^{-1}$
13:      $\alpha_{u,\mathrm{Im}}^t = (1 + (K^t\hat{d}_u^2(v_{\mathbf{z},\mathrm{Im}}^t - v_\mathbf{r}^t))^{-1})^{-1}$
14:      $x_u^{t+1} = \alpha_{u,\mathrm{Re}}^t\mathrm{Re}\{z_u^t\} + j\alpha_{u,\mathrm{Im}}^t\mathrm{Im}\{z_u^t\}$
15:      $\alpha_u^t = \alpha_{u,\mathrm{Re}}^t + \alpha_{u,\mathrm{Im}}^t$ (Onsager constant)
16:      $\rho_u^t = (2B/\beta)K^t\langle\hat{\mathbf{d}}^2\rangle\hat{d}_u^2$ (post-equalization SNR)
17:   **end for**
18:   $t = t + 1$
19: **end for**
20: **output:** L-MMSE estimate $\mathbf{z}^t$ and post-equalization SNR values ρ_u^t for each UE $u = 1, \ldots, U$

---

the imaginary part of $\mathbf{x}$ is zero. In fact, assuming a circularly-symmetric complex Gaussian prior for BPSK signals is a poor match as the imaginary part is zero. We generalize NOPE by estimating the signal variance E_x in (6) for the real and imaginary parts separately, which enables us to automatically adapt NOPE to the used constellation set. To do so, we decompose the weighted-norm of $\mathbf{z}^t$ denoted by $v_\mathbf{z}^t$, into real and imaginary parts, i.e., $v_\mathbf{z}^t = v_{\mathbf{z},\mathrm{Re}}^t + v_{\mathbf{z},\mathrm{Im}}^t$. More specifically, we can estimate the necessary variances as

$$\hat{E}_{\mathrm{x},\mathrm{Re}}^t = \frac{v_{\mathbf{z},\mathrm{Re}}^t - v_\mathbf{r}^t}{\sum_{u=1}^{U}\hat{d}_u^2} \quad \text{and} \quad \hat{E}_{\mathrm{x},\mathrm{Im}}^t = \frac{v_{\mathbf{z},\mathrm{Im}}^t - v_\mathbf{r}^t}{\sum_{u=1}^{U}\hat{d}_u^2},$$

for which $\hat{E}_{\mathrm{x},\mathrm{Re}}^t + \hat{E}_{\mathrm{x},\mathrm{Im}}^t = \hat{E}_\mathrm{x}^t$. With all these ingredients, we arrive at the generalized NOPE algorithm in Algorithm 2.

C. Numerical Results

Figure 1 shows uncoded bit error-rate (BER) simulation results in a $B = 64$ BS antenna, $U = 16$ UE massive MU-MIMO system with BPSK, 16-QAM, and 256-QAM. We show the performance of exact L-MMSE equalization, as well as the performance of NOPE for both infinite and fixed-precision. Solid lines correspond to floating-point precision, and circle markers correspond to fixed-point precision simulations of NOPE. Evidently, the BER performance of NOPE with $t_\mathrm{max} = 5$ iterations ($t_\mathrm{max} = 7$ for 256-QAM) is virtually indistinguishable from the exact L-MMSE estimator, which requires accurate knowledge of both the signal and noise powers. Due to its parameter free nature, NOPE is suitable for situations in which the signal and noise powers change rapidly (e.g., due to interference) or if the transmit constellation is unknown and must be estimated prior to data detection.
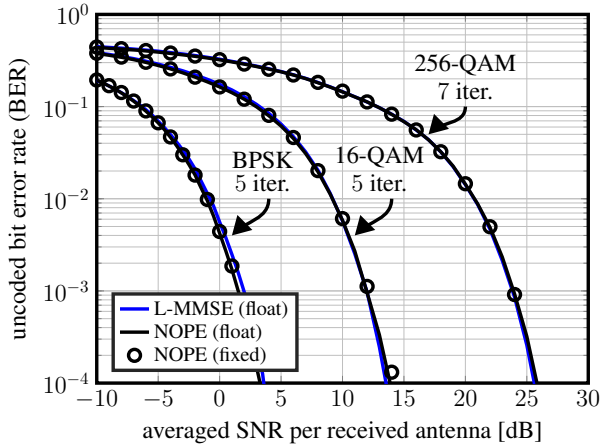
Fig. 1. Uncoded bit error-rate of NOPE algorithm in a 64×16 massive MU-MIMO system with Rayleigh fading channel matrices. NOPE closely approaches the performance of the L-MMSE estimator that requires exact knowledge of the signal and noise powers. Furthermore, fixed-point arithmetic in NOPE (shown with circle markers) does not exhibit a significant BER performance loss compared to infinite-precision arithmetic (continuous lines).
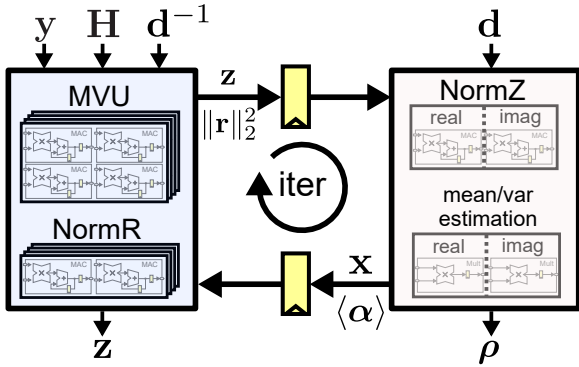

Fig. 2. Each iteration of the NOPE is partitioned into two main units: matrix-vector unit (left) and parameter estimation unit (right). Both units require an identical number of clock cycles, which allows us to process two independent problems via coarse-grained pipeline interleaving.

IV. VLSI ARCHITECTURE AND SYNTHESIS RESULTS

We now propose a very-large scale integration (VLSI) architecture of the NOPE algorithm for a $B = 64$ BS antenna, $U = 16$ UE massive MU-MIMO system. We then discuss the most essential optimization steps and finally present implementation results in a $28\,\mathrm{nm}$ CMOS technology.

A. Architecture Overview

We partition the NOPE iterations into two phases, each executed by a separate unit; see Fig. 2 for an architecture overview. The *matrix-vector unit* (MVU) executes the necessary matrix-vector multiplications and the *estimation unit* (EU) implements automatic parameter tuning.

The MVU performs the matrix-vector multiplication required to compute the 16 dimensional output vector $\mathbf{z}$ (line 7 of Alg. 2) and the scalar $\|\mathbf{r}\|_2^2$ (line 5 of Alg. 2). The EU implements the mean and variance estimation to compute the posterior mean $\mathbf{x}$ (line 14 of Alg. 2) and Onsager constant $\langle \boldsymbol{\alpha} \rangle$ (line 15 of Alg. 2). In addition, we compute the per-user post-equalization SNR $\boldsymbol{\rho}$ (line 16 of Alg. 2), which is required for log-likelihood
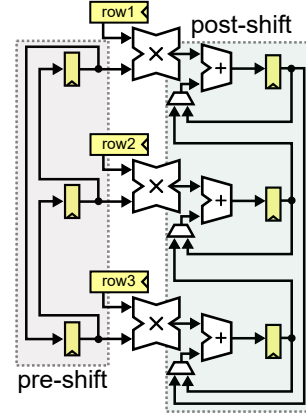

Fig. 3. Illustration of the matrix-vector unit (MVU) that computes computes $\mathbf{Hx}$ and $\mathbf{H}^{\mathrm{H}}\mathbf{r}$ in a 3×3 system via Cannon's algorithm. To compute $\mathbf{Hx}$, we circularly shift the inputs (pre-shift); to compute $\mathbf{H}^{\mathrm{H}}\mathbf{x}$, we circularly shift the outputs (post-shift). This approach enables column-wise storage of the entries of $\mathbf{H}$ without causing access contentions, leading to high throughput.

ratio (LLR) value calculations to perform soft-output data detection. Both units carry out their tasks in the same number of clock cycles, which enables us to process two independent equalization problems concurrently in the same architecture by means of coarse-grained pipeline interleaving.

B. Architecture Details

We now provide architecture details for the MVU and EU, and briefly discuss the key fixed-point implementation aspects.

1) MVU Details: The MVU computes both $\mathbf{Hx}$ and $\mathbf{H}^{\mathrm{H}}\mathbf{r}$ in a single unified architecture, similarly to the architecture in [8]. We divide the 64×16-dimensional channel matrix $\mathbf{H}$ into four 16×16 blocks, each of which are processed using a separate MVU, which we refer to as MVU-m, $m \in \{1,2,3,4\}$. Each 16×16 matrix-vector multiplication is carried out using 16 complex-valued multiply-accumulate (MAC) units; the matrix-vector operation is carried out on a column-by-column basis so that each MAC unit is associated to a row of the matrix.

A straightforward approach to compute $\mathbf{Hx}$ would be to broadcast the 16-dimensional vector $\mathbf{x}$ to all MVUs. To compute $\mathbf{H}^{\mathrm{H}}\mathbf{r}$ within the same architecture, access contentions would arise as one would need to be able to read all entries from the row of $\mathbf{H}^{\mathrm{H}}$ and sum all partial products. To enable highly parallel matrix-vector computation without causing access contentions, we use an architecture as depicted in Fig. 3 that performs a variant of Cannon's algorithm [17]. Let $\mathbf{A}$ be a 16×16 block of $\mathbf{H}$ where each row r is cyclically shifted by its index. To compute $\mathbf{Ax}$, the input $\mathbf{x}$ is first loaded into the input shift registers (the pre-shift block). We then circularly shift the entries of this shift register while sequentially calculating the MAC operations with entries of the matrix $\mathbf{A}$; the outputs are accumulated in the registers at the output of each MAC unit. This effectively implements a column-by-column matrix-vector operation in 16 clock cycles. To compute $\mathbf{H}^{\mathrm{H}}\mathbf{r}$, we load $\mathbf{r}$ into the input shift register but no cyclical shifts are carried out. Instead we cyclically exchange the outputs (the post-shift block) while accumulating the results. This effectively implements a row-by-row matrix-vector operation in 16 clock cycles.

TABLE I
SYNTHESIS RESULTS OF NOPE FOR A 64 BS ANTENNA, 16 UE SYSTEM
AND COMPARISON TO EXISTING MASSIVE MU-MIMO DATA DETECTORS.

	This work	Prabhu [9]	Tang [10]	Peng [11]	Castañeda [23]
System ($B \times U$)	64×16	128×8	128×32	128×8	128×8
Algorithm	NOPE	MMSE/ZF^a	MPD	MMSE	SDR
Parameters	none	E_x, N_0^{c}	E_x, N_0	E_x, N_0	E_x, N_0
Modulation	256-QAM	256-QAM	256-QAM	64-QAM	QPSK
Preproc. included	no	yes	no	yes	no
Preproc. quantities	col. gains	–	Gram mat.	–	Gram mat.
Results	synthesis	ASIC	ASIC	ASIC	post-layout
Technology [nm]	28	28	40	65	45
Area [mm^2]	0.28	1.10	0.58	2.57	0.48
Frequency [MHz]	800	300	425	680	560
Throughput [Gb/s]	0.92	0.30	2.76	1.02	0.13
Eff.^{b} [Gb/s/mm^2]	3.29	0.27	13.87	4.96	1.08

^{a}this design also supports precoding; ^{b}standard technology scaling rules apply; ^{c}the ZF mode does not require any parameters.

After the computation of $\mathbf{H}^H \mathbf{r}$, we have to accumulate the results of the four 16×16 blocks. We do this over two additional clock cycles: in cycle 1, MVU-1 and MVU-4 pass their result to MVU-2 and MVU-3 for accumulation; in cycle 2, MVU-2 passes its result to MVU-3 to obtain the final result.

2) EU Details: The EU computes the posterior mean $\mathbf{x}$ and Onsager constant $\langle \boldsymbol{\alpha} \rangle$. To this end, the EU first computes the 16-dimensional norm of the real and imaginary part of $\mathbf{z}$, $v_{\mathbf{z},\text{Re}}$ and $v_{\mathbf{z},\text{Im}}$. We employ two MAC units which compute the real and imaginary parts over 16 clock cycles. Once $v_{\mathbf{z},\text{Re}}$ and $v_{\mathbf{z},\text{Im}}$ are completed, we compute the so-called denoising parameter $\alpha_{u,\text{Re}}$ (line 12 of Alg. 2) and $\alpha_{u,\text{Im}}$ (line 13 of Alg. 2) for each u-th UE sequentially over 16 clock cycles. We note that the function $(1 + x^{-1})^{-1} = 1 - (1 + x)^{-1} \in [0, 1)$ in lines 12 and 13 of Alg. 2 is numerically stable so we employ a single-iteration of the LUT-based Newton-Raphson procedure.

3) Fixed-Point Arithmetic: In order to achieve low hardware complexity and high throughput, our design uses fixed-point arithmetic. We first globally scale $\mathbf{H}$ so that the real and imaginary entries are less than 1. We then quantize each element of $\mathbf{H}$ to 10 fraction bits, and $\mathbf{y}$ to 6 integer and 4 fraction bits. The fixed-point performance of our NOPE design is shown in Fig. 1. The solid lines correspond to floating-point performance, the markers to the fixed-point performance of our golden model.

C. Implementation Results and Conclusion

Table I shows synthesis results for NOPE in a 28 nm CMOS technology and compares our design to existing data detectors for massive MU-MIMO. We note that the numbers reported in Table I for NOPE are based on synthesis results; an ASIC design is part of ongoing work. While our design is comparable to other designs in terms of hardware efficiency, we emphasize that NOPE is completely parameter-free (other than knowledge of $\mathbf{H}$ and $\mathbf{y}$), which makes it more resilient to parameter mismatch and dynamic variations of the system compared to all the other methods. In addition, NOPE requires a minimal amount of preprocessing, i.e., $\hat{\mathbf{d}}^2$ and $\hat{\mathbf{d}}^{-2}$, in contrast to, e.g., the design of [10] that requires computation of the Gram which often dominates the complexity of massive MU-MIMO data detectors [3]. In summary, NOPE is a robust "fire-and-forget"

equalization algorithm for MU-MIMO systems that achieves L-MMSE performance at competitive implementation complexity.

REFERENCES

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

[2] J. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. Soong, and J. Zhang, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[3] 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.

[4] S. Verdú, *Multiuser Detection*, 1st ed.   Cambridge Univ. Press, 1998.

[5] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO: How many antennas do we need?" in *Proc. Allerton Conf. Commun., Contr., Comput.*, Sept. 2011, pp. 545–550.

[6] R. Ghods, C. Jeon, G. Mirza, A. Maleki, and C. Studer, "Optimally-tuned nonparametric linear equalization for massive MU-MIMO systems," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 2118–2122.

[7] C. Studer, S. Fateh, and D. Seethaler, "ASIC implementation of soft-input soft-output MIMO detection using MMSE parallel interference cancellation," *IEEE J. Solid-State Circuits*, vol. 46, no. 7, pp. 1754–1765, Jul. 2011.

[8] O. Castañeda, S. Jacobsson, G. Durisi, M. Coldrey, T. Goldstein, and C. Studer, "1-bit massive MU-MIMO precoding in VLSI," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, 2017.

[9] H. Prabhu, J. N. Rodrigues, L. Liu, and O. Edfors, "3.6A 60pJ/b 300Mb/s 128×8 Massive MIMO precoder-detector in 28nm FD-SOI," in *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, Feb 2017, pp. 60–61.

[10] W. Tang, C. H. Chen, and Z. Zhang, "A 0.58mm^2 2.76Gb/s 79.8pJ/b 256-QAM massive MIMO message-passing detector," in *IEEE Symp. VLSI Circuits*, Jun. 2016, pp. 1–2.

[11] G. Peng, L. Liu, S. Zhou, S. Yin, and S. Wei, "A 1.58 Gbps/W 0.40 Gbps/mm^2 ASIC implementation of MMSE detection for 128×8 64-QAM Massive MIMO in 65 nm CMOS," *IEEE Trans. Circuits Syst. I*, 2017.

[12] C. Studer, M. Wenk, and A. Burg, "MIMO transmission with residual transmit-RF impairments," in *Int. ITG Workshop on Smart Antennas (WSA)*, Feb. 2010, pp. 189–196.

[13] E. Björnson, J. Hoydis, M. Kountouris, and M. Debbah, "Massive MIMO systems with non-ideal hardware: Energy efficiency, estimation, and capacity limits," *IEEE Trans. Inf. Theory*, vol. 60, no. 11, pp. 7112–7139, Apr. 2014.

[14] D. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Natl. Academy of Sciences (PNAS)*, vol. 106, no. 45, pp. 18 914–18 919, Sept. 2009.

[15] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *The Annals of Statistics*, vol. 9, no. 6, pp. 1135–1151, Nov. 1981.

[16] C. Jeon, A. Maleki, and C. Studer, "On the performance of mismatched data detection in large MIMO systems," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2016, pp. 180–184.

[17] L. Cannon, "A cellular computer to implement the Kalman filter algorithm," Ph.D. dissertation, Montana State University, USA, 1969.

[18] U. Madhow and M. L. Honig, "MMSE interference suppression for direct-sequence spread-spectrum CDMA," *IEEE Trans. Commun.*, vol. 42, no. 12, pp. 3178–3188, Aug. 1994.

[19] K. R. Kumar, G. Caire, and A. L. Moustakas, "Asymptotic performance of linear receivers in MIMO fading channels," *IEEE Trans. Inf. Theory*, vol. 55, no. 10, pp. 4398–4418, Oct. 2009.

[20] J. Hoydis, S. Ten Brink, and M. Debbah, "Massive MIMO: How many antennas do we need?" in *Allerton Conf. on Commun., Contr, and Comput.* IEEE, 2011, pp. 545–550.

[21] D. Tse and S. Hanly, "Linear multiuser receivers: effective interference, effective bandwidth and user capacity," *IEEE Trans. Inf. Theory*, vol. 45, no. 2, pp. 641–657, Mar. 1999.

[22] C. D. Perels, "Frame-based MIMO-OFDM systems: impairment estimation and compensation," Ph.D. dissertation, ETH Zurich, Switzerland, 2007.

[23] O. Castañeda, T. Goldstein, and C. Studer, "Data detection in large multi-antenna wireless systems via approximate semidefinite relaxation," *IEEE Trans. Circuits Syst. I*, vol. 63, no. 12, pp. 2334–2346, Dec. 2016.