Efficient Expectation Propagation Massive MIMO Detector with Neumann-Series Approximation

Xiaosi Tan, Weihong Xu, Yaping Zhang, Xiaohu You, *Fellow, IEEE*, and Chuan Zhang, *Member, IEEE*

***Abstract*—Expectation propagation (EP) attains near-optimal performance for massive multiple-input multiple-output (MIMO) detection. However, the inevitable matrix inversions and exponentiations at each EP iteration bring great challenges to realistic hardware implementation. To address these issues, a low-complexity EP with iterative Neumann-series Approximation (EP-INSA) detector is proposed by employing INSA to estimate the inverse matrices. Further approximations are applied to avoid the exponentiations, which makes EP-INSA an efficient, feasible, and hardware-friendly detector for massive MIMO with various modulations. Simulation results show that EP-INSA attains similar performance as exact EP with only a few INSA terms, which assures enhanced performance and complexity trade-off. The associated hardware architectures of EP-INSA detector are also presented. The implementation results on 65 nm CMOS technology show that our design yields a throughput of 0.62 Gb/s with $2.63\times$ area efficiency of existing EP detector.**

***Index Terms*—Massive MIMO, expectation propagation, iterative Neumann-series Approximation, hardware implementation**

I. Introduction

MASSIVE multiple-input multiple-output (MIMO) is widely believed to be one of the key technologies for 5G wireless systems [1]. However, the enormous antenna array imposes great challenges including high complexity and unsatisfactory performance on signal detection problems [2]. Conventional detectors for small-scale MIMO are no longer practical for large-scale systems. Optimal algorithms like sphere decoding (SD) [3] suffer prohibitive computational costs. Linear detectors as minimum mean square error (MMSE) [4] are simpler to implement, but attain unsatisfying performances.

Recently, iterative message passing detectors (MPDs) based on Bayesian inference are broadly researched. It is shown that MPDs such as belief propagation (BP) [5], approximate message passing (AMP) [6], and channel-hardening exploiting message passing (CHEMP) [7] achieve superior performance with low complexity compared to linear algorithms. Nevertheless, they still suffer great performance loss with certain MIMO configurations [8].

Expectation propagation (EP) [9] is firstly considered for MIMO detection in [10]. The proposed iterative EP detector (EPD) constructs a Gaussian approximation for the posterior distribution of the transmitted symbols. As shown in [10], EPD achieves promising near-optimal performance against various antenna configurations and modulations, which outperforms sate-of-the-art (SOA) detectors including MMSE, BP, and Gaussian tree approximation (GTA) [8]. However, each iteration of EPD involves an inevitable matrix inversion whose complexity is $\mathcal{O}(M^3)$ (M denotes the number of users in MIMO systems). As the system size grows larger, EPD suffers unaffordable computational pressure. To relieve this complexity burden, approximations of EP are proposed. For example, in [11], the matrix inversions are substituted by Neumann-series approximation (NSA), which forms the EP-NSA detector. EP-successive updating (EP-SU) is proposed in [12] which replaces the inversions by Sherman–Morrison formula. However, computational savings attained by these methods are limited since a complexity of $\mathcal{O}(M^3)$ is still required to guarantee satisfying performance.

Contributions: In this brief, an efficient EP with iterative NSA (EP-INSA) massive MIMO detector is proposed, in which the matrix inversions are approximated by INSA to achieve overall $\mathcal{O}(M^2)$ complexity. Moreover, approximation schemes are adopted to avoid exponentiations, which makes the algorithm hardware-friendly. Numerical results show that EP-INSA maintains similar good performance as EPD with reduced complexity. The efficient and flexible hardware architectures for EP-INSA are also developed to accommodate different antenna configurations and modulations for massive MIMO systems.The post-layout results on 65 nm 1P9M CMOS technology are also presented to demonstrate the advantages of EP-INSA algorithm.

Brief Outline: The remainder of this brief is organized as follows. Section II introduces the MIMO system model and EPD. EP-INSA detector is proposed in Section III with detailed descriptions of employed approximations. Section IV exhibits numerical results. The corresponding hardware architecture is depicted in Section V with implementation results and comparisons with SOAs. Section VI concludes the brief.

Notations: Lower- and upper-case boldface letters stand for column vector and matrix, respectively. The i-th element in vector $\mathbf{x}$ is denoted by x_i, while the (i,j)-th element in matrix $\mathbf{X}$ is X_{ij}. $\mathbf{I}_N$ indicates an $N\times N$ identity matrix. $\mathcal{N}(\mathbf{y}:\boldsymbol{\mu},\boldsymbol{\Sigma})$ denotes the Gaussian function with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. $diag(\mathbf{x})$ creates a matrix with $\mathbf{x}$ in the diagonal, and $diag(\mathbf{X})$ is the vector of the diagonal elements of $\mathbf{X}$.

II. Preliminaries

A. System Model

A massive MIMO system equipped with N receiving antennas at the base station (BS) serving M single-antenna

Xiaosi Tan, Weihong Xu, Yaping Zhang, Xiaohu You, and Chuan Zhang are with LEADS, the National Mobile Communications Research Laboratory of Southeast University, and the Purple Mountain Laboratories Nanjing, China. Email: chzhang@seu.edu.cn. (*Corresponding author: Chuan Zhang.*)

This work was presented in part at IEEE International Symposium on Circuits and Systems (ISCAS) Late Breaking News (LBN), Sapporo, Hokkaido, Japan, 2019.

users is considered (usually $M < N$). Suppose that the transmitted vector is denoted by $\mathbf{x} \in \Theta^{M\times 1}$, where Θ is the modulation alphabet. The received vector $\mathbf{y} \in \mathbb{C}^{N\times 1}$ is obtained by the system model:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}, \tag{1}$$

where $\mathbf{H} \in \mathbb{C}^{N\times M}$ is the channel matrix for i.i.d Rayleigh-fading channels, $\mathbf{n} \in \mathbb{C}^{N\times 1}$ is the additive *Gaussian* white noise (AWGN) vector with mean zero and variance σ_n^2. $\mathbf{H}$ is supposed to be known at the receiver side in this brief. Let $\mathbb{I}_{x_i\in\Theta}$ be the indicator function. According to [5], the posterior distribution of $\mathbf{x}$ can be expressed as

$$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x})f(\mathbf{x}) \propto \mathcal{N}(\mathbf{y} : \mathbf{H}\mathbf{x}, \sigma_n^2\mathbf{I}_N)\prod_{i=1}^{M}\mathbb{I}_{x_i\in\Theta}. \tag{2}$$

B. Expectation Propagation for MIMO Detection

EP [9] is a Bayesian learning technique which constructs an exponential family distribution $q(\mathbf{x})$ to approximate the intractable posterior $p(\mathbf{x})$. The EPD proposed in [10] updates $q(\mathbf{x})$ iteratively by forcing moment matching conditions [9]. Specifically, we map Eq. (1) to the real domain by real value decomposition [5]. Then the desired exponential approximation $q(\mathbf{x})$ of Eq. (2) can be formulated by substituting the non-Gaussian parts in Eq. (2) with exponential distributions as:

$$q(\mathbf{x}) \propto \mathcal{N}(\mathbf{y} : \mathbf{H}\mathbf{x}, \sigma_n^2\mathbf{I}_{2N})\exp(-\frac{1}{2}\mathbf{x}^T\mathbf{\Lambda}\mathbf{x} + \boldsymbol{\gamma}^T\mathbf{x}), \tag{3}$$

where $\boldsymbol{\gamma} \in \mathbb{R}^{2N}$, and $\mathbf{\Lambda} = diag([\Lambda_1, \Lambda_2, \ldots, \Lambda_{2N}])$. As shown in [10], the mean $\boldsymbol{\mu}$ and variance $\mathbf{\Sigma}$ of $q(\mathbf{x})$ can be computed by:

$$\mathbf{\Sigma} = (\frac{1}{\sigma_n^2}\mathbf{H}^T\mathbf{H} + \mathbf{\Lambda})^{-1}, \tag{4}$$

$$\boldsymbol{\mu} = \mathbf{\Sigma}(\frac{1}{\sigma_n^2}\mathbf{H}^T\mathbf{y} + \boldsymbol{\gamma}). \tag{5}$$

In EPD, the parameters (γ_i, Λ_i) are updated through the following iterative process for $i = 1, \ldots, 2M$. At the l-th iteration with the input $(\boldsymbol{\gamma}^{(l-1)}, \mathbf{\Lambda}^{(l-1)})$:

1) Update $\mathbf{\Sigma}^{(l)}$ and $\boldsymbol{\mu}^{(l)}$ by Eq.s (4) and (5). The mean and variance of the i-th marginal $q^{(l)}(x_i)$ can then be expressed as $\mu_i^{(l)}$ and $\sigma_i^{2(l)} = \Sigma_{ii}^{(l)}$, respectively.
2) Update parameters $(t_i^{(l)}, h_i^{2(l)})$ in the cavity marginal

$$q^{(l)\backslash i}(x_i) = \frac{q^{(l)}(x_i)}{\exp(-\frac{1}{2}\Lambda_i^{(l)}x_i^2 + \gamma_i^{(l)}x_i)} \propto \mathcal{N}(t_i^{(l)}, h_i^{2(l)}); \tag{6}$$

3) Compute the mean η_i and the variance χ_i of the i-th marginal distribution of $\hat{p}^{(l)}(\mathbf{x})$

$$\hat{p}^{(l)}(x_i) = q^{(l)\backslash i}(x_i)\mathbb{I}_{x_i\in\Theta}. \tag{7}$$

4) Update $(\gamma_i^{(l+1)}, \Lambda_i^{(l+1)})$ by respectively matching $\eta_i^{(l)}$ and $\chi_i^{(l)}$ with the mean and variance of the distribution

$$\tilde{q}^{(l+1)}(x_i) = q^{(l)\backslash i}(x_i)\exp(-\frac{1}{2}\Lambda_i^{(l+1)}x_i^2 + \gamma_i^{(l+1)}x_i) \tag{8}$$

.

Details of the algorithm can be found in [10]. From above it's obvious that the computational complexity and implementation challenges of EPD is mostly dominated by two main steps: (1) The update of $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ in Eq.s (4) and (5) which involves a matrix inversion of complexity $\mathcal{O}(M^3)$; (2) The computation of mean and variance of Eq. (7) which requires multiple exponentiations and divisions with higher-order modulations.

III. The Proposed EP-INSA Detector

According to [11], the matrix inversion in Eq. (4) can be approximated by NSA to alleviate the computational burden. Specifically, let $\mathbf{W} = \sigma_n^{-2}\mathbf{H}^T\mathbf{H} + \mathbf{\Lambda}$ and $\mathbf{D}$ be the main diagonal matrix of $\mathbf{W}$. The inversion of $\mathbf{W}$, or equivalently $\mathbf{\Sigma}$, can be estimated by the k-term NSA:

$$\mathbf{\Sigma} \approx \mathbf{W}_k^{-1} \approx \sum_{n=0}^{k-1}(-\mathbf{D}^{-1}\mathbf{E})^n\mathbf{D}^{-1}, \tag{9}$$

where $\mathbf{E} = \mathbf{W} - \mathbf{D}$. Replacing Eq. (4) by Eq. (9), the EP-NSA algorithm is constructed. However, to guarantee satisfying bit error rate (BER) performance, it's required that $k \geq 3$ in Eq. (9) for most antenna configurations as presented in [11], which still demands a computational complexity of $\mathcal{O}(M^3)$.

To enhance the efficiency, we propose the INSA scheme to approximate Eq. (4) while updating Eq. (5) iteratively at the same time. Let $\Psi = -\mathbf{D}^{-1}\mathbf{E} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{W}$, Eq. (9) becomes

$$\mathbf{W}_k^{-1} = \sum_{n=0}^{k-1}\Psi^n\mathbf{D}^{-1} = \Psi\mathbf{W}_{k-1}^{-1} + \mathbf{D}^{-1}. \tag{10}$$

The above Eq. (10) provides a sequential scheme to compute the k-term NSA $\mathbf{W}_k^{-1}$. Meanwhile, notice that only the diagonal elements of $\mathbf{\Sigma}$, i.e., σ_i^2's, are required in the EP iterations. Therefore, we only update the diagonal elements in the proposed INSA iterations. Specifically, at the i-th iteration for computing $\mathbf{W}_k^{-1}$ $(2 \leq i \leq k)$,

$$diag(\mathbf{W}_k^{-1(i)}) = diag(\Psi\mathbf{W}_k^{-1(i-1)}) + diag(\mathbf{D}^{-1}). \tag{11}$$

Denote $\mathbf{b} = \sigma_n^{-2}\mathbf{H}^T\mathbf{y} + \boldsymbol{\gamma}$. From Eq. (5) we can deduct that $\boldsymbol{\mu} = \mathbf{\Sigma}\mathbf{b} \approx \mathbf{W}_k^{-1}\mathbf{b}$. Multiplying Eq. (10) with $\mathbf{b}$, an iterative approximation of $\boldsymbol{\mu}$ can also be formulated, where at the i-th iteration $(2 \leq i \leq k)$ we have

$$\boldsymbol{\mu}^{(i)} = \Psi\boldsymbol{\mu}^{(i-1)} + \mathbf{D}^{-1}\mathbf{b}. \tag{12}$$

The proposed INSA algorithm to update $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ is summarized in Alg. 1. All the computations of Eq.s (5) and (4) in EPD can be replaced with this INSA procedure.

Moreover, to relief the computational burden in Step 3 of the EPD, we introduce the simplification schemes following [13]. Indeed, we need to calculate the $1st$ and $2nd$ moments to get the mean and variance of Eq. (7). The exact m-th moment has the form $\sum_{x_i\in\Theta} x_i^m\mathcal{N}(x_i|t_i, h_i^2)/\sum_{x_i\in\Theta}\mathcal{N}(x_i|t_i, h_i^2)$. Let $|\Theta|$ denote the constellation size. Each EP iteration then involves $M|\Theta|$ exponentiations, which are unaffordable to implement for higher-order modulations. In this work, the mean of Eq. (7) is approximated with a hard decision on t_i, while the variance is estimated by a linear function of t and h^2 following [13] with real coefficients P_0, P_1, and P_2. Combining

Algorithm 1: INSA for updating $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ in EP

Input: $\mathbf{W}$, $\mathbf{b}$
function INSAupdate($\mathbf{W}, \mathbf{b}, k$)
  $\mathbf{D} = \text{diag}(\mathbf{W})$, $\Psi = \mathbf{I} - \mathbf{D}^{-1}\mathbf{W}$;
  $\boldsymbol{\mu}^{(1)} = \mathbf{D}^{-1}\mathbf{W}$, $\mathbf{W}^{-1(1)} = \mathbf{D}^{-1}$;
  for $i = 2, \ldots, k$ **do**
    $\boldsymbol{\mu}^{(i)} = \Psi\boldsymbol{\mu}^{(i-1)} + \boldsymbol{\mu}^{(1)}$;
    $diag(\mathbf{W}^{-1(i)}) =$
     $diag(\Psi\mathbf{W}^{-1(i-1)}) + diag(\mathbf{D}^{-1})$.
  end
end
Output: $\boldsymbol{\mu} = \boldsymbol{\mu}^{(k)}$, $\boldsymbol{\sigma}^2 = \text{diag}(\mathbf{W}^{-1(k)})$

all the above approximations, the proposed EP-INSA algorithm is summarized in Alg. 2. Message damping is applied in this algorithm with damping factors α and β.

Algorithm 2: Proposed EP-INSA Algorithm

Input: $\mathbf{A} = \sigma_n^{-1}\mathbf{H}^T\mathbf{H}$, $\tilde{\mathbf{b}} = \sigma_n^{-1}\mathbf{H}^T\mathbf{y}$, α, β, L
$\Lambda_i^{(0)} = 0$, $\gamma_i^{(0)} = E_s^{-1}$, $i = 1, \ldots, 2M$;
$\mathbf{W}^{(0)} = \mathbf{A} + \text{diag}(\boldsymbol{\Lambda}^{(0)})$, $\mathbf{b}^{(0)} = \tilde{\mathbf{b}} + \boldsymbol{\gamma}^{(0)}$;
$\boldsymbol{\mu}^{(0)}, \boldsymbol{\sigma}^{2(0)} = \text{INSAupdate}(\mathbf{W}^{(0)}, \mathbf{b}^{(0)}, k)$;
for $l = 1, \ldots, L$ **do**
  for $i = 1, \ldots, 2M$ *(or parallel execution)* **do**
    $h_i^{2(l)} = \frac{1}{\sigma_i^{-2(l)} - \Lambda_i^{(l)}}$; $t_i^{(l)} = h_i^{2(l)}(\frac{\mu_i^{(l)}}{\sigma_i^{-2(l)}} - \gamma_i^{(l)})$;
    $\hat{\mu}_i^{(l)} = \text{Hard\_Decision}(t_i^{(l)})$;
    $\hat{\sigma}_i^{2(l)} = P_0 + P_1 t_i^{(l)} + P_2 h_i^{2(l)}$;
    $\Lambda_i^{(l)} = \frac{1}{\hat{\sigma}_i^{2(l)}} - \frac{1}{h_i^{2(l)}}$; $\gamma_i^{(l)} = \frac{\hat{\mu}_i^{(l)}}{\hat{\sigma}_i^{2(l)}} - \frac{t_i^{(l)}}{h_i^{2(l)}}$;
    $\Lambda_i^{(l)} = \alpha\Lambda_i^{(l)} + (1-\alpha)\Lambda_i^{(l-1)}$;
    $\gamma_i^{(l)} = \beta\gamma_i^{(l)} + (1-\beta)\gamma_i^{(l-1)}$;
    $\boldsymbol{\mu}^{(l)}, \boldsymbol{\sigma}^{2(l)} = \text{INSAupdate}(\mathbf{W}^{(l)}, \mathbf{b}^{(l)}, k)$
  end
  Output: Hard_Decision($\boldsymbol{\mu}^{(L)}$)
end

The complexity of the INSA approximation in Alg. 1 is $\mathcal{O}(kM^2)$. Also, the approximation for mean and variance of Eq. (7) makes the complexity of Alg. 2 independent of the modulation order. Therefore, the overall complexity of EP-INSA is $\mathcal{O}(kM^2L + ML)$, where L denotes the number of iterations. Compared to the complexities of MMSE, exact EPD, and EP-NSA ($k > 2$) which are all at the $\mathcal{O}(M^3)$ level, EP-INSA achieves considerable computational savings.

IV. Simulation Results

The BER performance of the proposed EP-INSA is investigated and compared with MMSE and EPD [10] as shown in Fig. 1. An uncoded MIMO system is considered with fixed $N = 128$ and various M's, while three modulations, 16-QAM, 64-QAM, and 256-QAM, are employed for comparison. Fig. 1a shows the BER performance with $M = 8$, in which EP-INSA with $k = 2, 3$ and 4 are presented. It can be observed that $k = 3$ is sufficient for EP-INSA to obtain similar performance as the EPD for all modulations, which outperforms MMSE. When $M = 16$ as exhibited in Fig. 1b, we need $k = 4$ to ensure that EP-INSA can recover the performance of EPD. For $M = 32$ in Fig. 1c, k has to be raised to 5 for EP-INSA to attain comparable BER as EPD with slight degradation. For example, at the BER of 10^{-5} for 256-QAM, EP-INSA has an about 0.4 dB performance loss compared to EPD, but meanwhile outperforms MMSE for 1 dB. Overall, EP-INSA can achieve similar BER performance as exact EPD with only a few terms in INSA. For MIMO systems with a certain N, larger M will require greater k and a few more iterations for EP-INSA to attain comparable BER as the exact EP. However, considering the reduced complexity, the proposed EP-INSA still attains improved efficiency compared to EPD.

V. Hardware Implementation

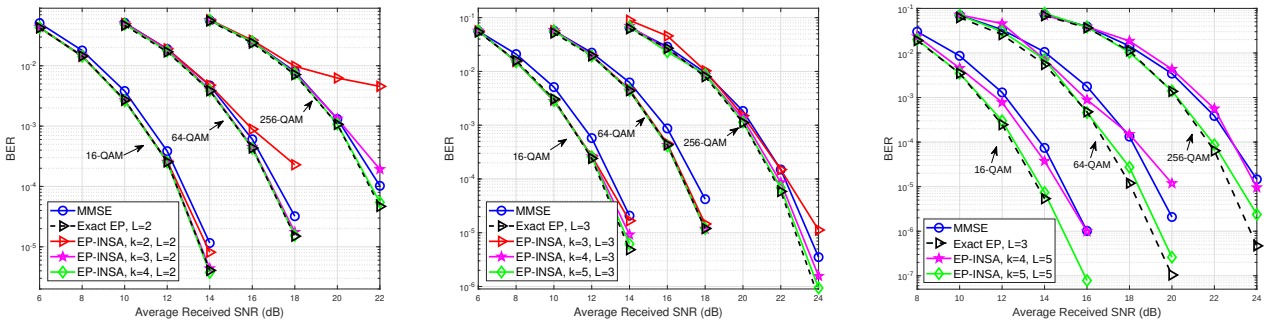
A. Hardware Architecture

1) Overall Architecture: Fig. 2 shows the overall diagram of the proposed EP-INSA detector, which mainly contains three modules, namely the preprocessing unit, the Neumann unit, and approximate moment matching unit. First, $\mathbf{H}$ and $\mathbf{y}$ are preprocessed in the preprocessing unit to obtain the Gram matrix $\mathbf{H}^T\mathbf{H}$ and the matched filter output $\mathbf{H}^T\mathbf{y}$. Then EP-INSA is initialized in the Neumann unit. After that, the iteration process of EP-INSA follows which involves two steps including Neumann updating and approximate moment matching.

2) Systolic-based Preprocessing Unit: As illustrated in Fig. 3, the preprocessing unit is composed of matched filter and Gram matrix. Two types of processing element (PE), namely PE-A and PE-B, are implemented in this unit. Each PE contains a stage-pipelined complex multiplier and accumulator (MAC). Each PE-B accepts new input and delivers the latched data to its two adjacent PEs horizontally and vertically. In comparison, each PE-A only accepts new input horizontally and propagates the latched data vertically.

The Gram matrix is a diagonal-based systolic array consisting of total $\frac{M(M+1)}{2}$ PEs. Precisely, M PE-As are located in the diagonal of the systolic array while the rest $\frac{M(M-1)}{2}$ PE-Bs are located in the lower triangular part. Each row of $\mathbf{H}$ is fed into the systolic array sequentially. $M + N$ clock cycles are required to compute the Gram matrix. The M PE-Bs inside matched filter are cascaded into 1D pattern to calculate the correct results. Each PE receives the data of $\mathbf{H}$ and $\mathbf{y}$ from its left side and bottom, respectively. The incoming data is multiplied and accumulated in each PE. The matched filter results can be obtained after $M + N$ clock cycles.

3) Neumann Unit: The Neumann unit first calculates the approximate matrix inversion and then updates $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$. The corresponding block diagram is depicted in Fig. 4. The *add.*, *reci.*, and *mult.* modules contain $2M$ adders, reciprocal units, and real multipliers, respectively. The matrix vector multiplication contains a $2M \times 2M$ multiplier array and $2M$ adder trees to compute the matrix vector product. The proposed Neumann unit computes the approximate inversion with $\mathcal{O}(M^2)$ complexity through reusing the intermediate $\boldsymbol{\mu}^{(i)}$ as suggested in [14]. The reciprocal unit to calculate $\mathbf{D}^{-1}$ is based on the method in [15]. It is implemented by two multipliers and



(a) $M = 8$, $N = 128$ (b) $M = 16$, $N = 128$ (c) $M = 32$, $N = 128$

Fig. 1. The BER performance of proposed EP-INSA compared with MMSE and exact EPD [10].

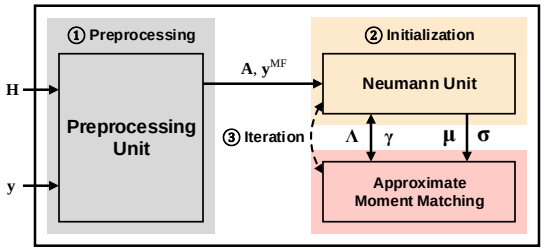


Fig. 2. Overall diagram of proposed approximate EP detector.

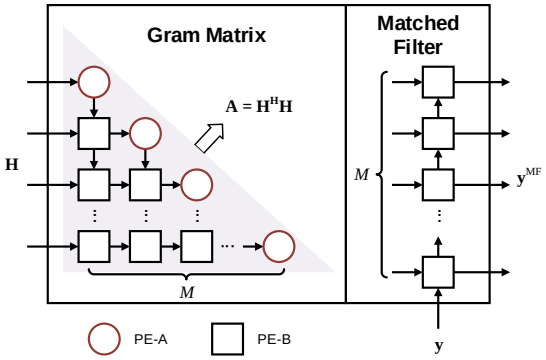


Fig. 3. Architectures of systolic-based processing unit.

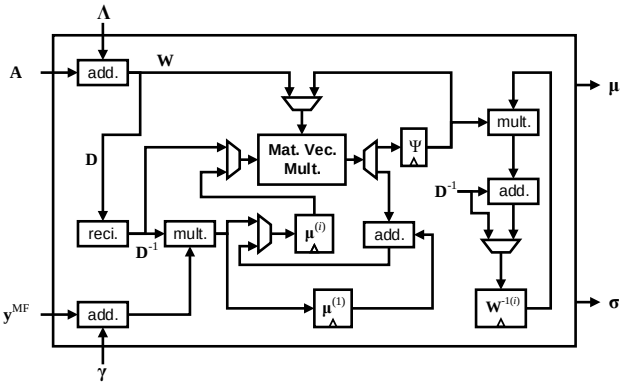


Fig. 4. Block diagram of Neumann unit.

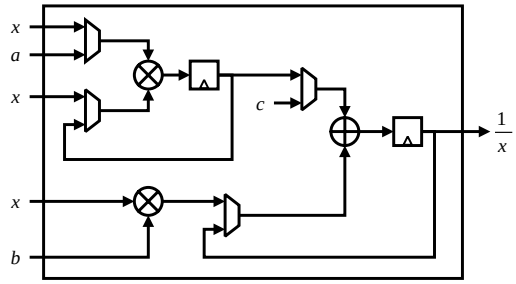


Fig. 5. Block diagram of reciprocal unit, where $a = 2.65548$, $b = -5.92781$, and $c = 4.28387$.

one adder as shown in Fig. 5. The coefficients are set to $a = 2.65548$, $b = -5.92781$, and $c = 4.28387$, which provide sufficient arithmetic precision for the INSA iteration.

4) Approximate Moment Matching Unit: The approximate moment matching unit receives $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ from the Neumann unit and then computes the updated $\boldsymbol{\Lambda}$ and $\boldsymbol{\gamma}$. As mentioned in Section III, the exact mean $\mu_i^{(l)}$ and variance $\sigma_i^{2(l)}$ calculations are estimated by the hard decision and polynomial function, respectively. As a result, the complexity of moment matching is significantly reduced.

B. Implementation Results and Comparison

The proposed EP detector with Neumann approximation is implemented on SMIC 65-nm 1P9M CMOS process and synthesized by Synopsys Design Compiler. The design is placed and routed using Synopsys IC Compiler. The power dissipation is estimated by Synopsys Prime Time PX.

The processing data of proposed EP-INSA is quantized to 14 bits as in [16], causing negligible performance degradation. The iteration number of Neumann unit is set to 3 for the initialization process and 2 for the EP iteration process. The EP detector is implemented with the antenna size of 8×128.

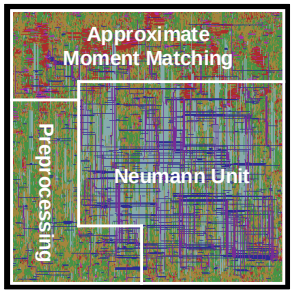
The layout photo and implementation results are illustrated in Fig. 6. The chip integrates a total of 1243k logic gates and runs at the frequency of 500 MHz with the supply voltage of 1.0 V. Under 256-QAM modulation, the proposed EP-INSA delivers a throughput of 0.62 Gb/s with 573.1 mW power dissipation. It should be noted that the proposed EP-INSA natively supports different modulations from QPSK to 256-QAM without modification to the hardware architectures.

The implementation results are compared with several SOAs

TABLE I
HARDWARE COMPARISON FOR DIFFERENT MIMO DETECTORS

Detector	This work	Peng [16] [TCAS-I, 2017]	Chen [17] [TCAS-I, 2018]	Tang [13] [ISSCC, 2018]
$M \times N$	8×128	8×128	8×128	16×128
Modulation	QPSK to 256-QAM	64-QAM	QPSK	QPSK to 256-QAM
Algorithm	EP-INSA	MMSE	MPD	EP
Preprocessing Unit	**Included**	**Included**	Not included	Not included
Process [nm]	65	65	40	28
Supply Voltage [V]	1.0	1.0	0.9	1.0
Gate Count [kGE]	1243	1070	1167	3607
Frequency [MHz]	500	680	500	512
Power [mW]	573.1	650	501	127
Throughput [Gb/s]	0.62	1.02	8.0	1.60
Scaled‡ **Area Eff.** [Gb/s/GE]	0.50	0.95	4.22	0.19
Scaled‡ **Energy Eff.** [pJ/b]	0.92	0.64	0.20	0.43

‡ Scaled to 65 nm and 1.0 V with frequency $\propto s$ and power $\propto \frac{1}{s}(\frac{1.0}{V_{dd}})^2$, where s is the scaling factor to 65 nm.
◇ Area efficiency = throughput / gate count and energy efficiency = power / throughput.



Technology	65 nm 1P9M CMOS
Antenna Size	8×128
Modulation	QPSK to 256-QAM
Gate Count (kGE)	1243
Supply Voltage (V)	1.0
Clock Freq. (MHz)	500
Power (mW)	573.1
Throughput (Gb/s)	0.62

Fig. 6. Layout and implementation results of EP-INSA detector.

[13, 16, 17] in Table I. The MIMO detectors in [16], [17], and [13] are based on MMSE, MPD, and exact EPD, respectively. Only the proposed EP-INSA and MMSE detector in [16] include the preprocessing unit. We observe that EP-INSA achieves higher area efficiency than the exact EPD in [13] due to the low-complexity INSA. It should be noted that the MPD [17] significantly reduces the complexity by exploiting the channel hardening technique [7]. However, the MPD can only guarantee near-MMSE performance under specific antenna configurations. In contrast, the BER performance of EP-based detector outperforms MMSE and is near-optimal under various channel conditions [13].

VI. CONCLUSION

In this brief, an efficient EP-INSA algorithm is proposed for massive MIMO detection, in which approximation schemes are employed to avoid the direct matrix inversion and exponentiations in EPD to achieve reduced complexity while guaranteeing good BER performance. Moreover, the proposed EP-INSA algorithm implemented on 65 nm technology obtains higher efficiency than existing SOA EP detector.

REFERENCES

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

[2] C. Zhang, Y. Huang, F. Sheikh, and Z. Wang, "Advanced Baseband Processing Algorithms, Circuits, and Implementations for 5G Communication," *IEEE J. Emerging Sel. Top. Circuits Syst.*, vol. 7, no. 4, pp. 477–490, Dec. 2017.

[3] L. G. Barbero and J. S. Thompson, "Fixing the complexity of the sphere decoder for MIMO detection," *IEEE Trans. Wireless Commun.*, vol. 7, no. 6, pp. 2131–2142, Jun. 2008.

[4] G. Caire, R. R. Muller, and T. Tanaka, "Iterative multiuser joint decoding: Optimal power allocation and low-complexity implementation," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 1950–1973, Sep. 2004.

[5] J. Yang, W. Song, S. Zhang, X. You, and C. Zhang, "Low-Complexity Belief Propagation Detection for Correlated Large-Scale MIMO Systems," *J. Sign. Process Syst.*, vol. 90, no. 4, pp. 585–599, Apr. 2018.

[6] C. Jeon, R. Ghods, A. Maleki, and C. Studer, "Optimality of large MIMO detection via approximate message passing," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2015, pp. 1227–1231.

[7] T. L. Narasimhan and A. Chockalingam, "Channel Hardening-Exploiting Message Passing (CHEMP) Receiver in Large-Scale MIMO Systems," *IEEE J. Sel. Topics Signal Process*, vol. 8, no. 5, pp. 847–860, Oct. 2014.

[8] J. Goldberger and A. Leshem, "MIMO Detection for High-Order QAM Based on a Gaussian Tree Approximation," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 4973–4982, Aug. 2011.

[9] T. P. Minka, "Expectation Propagation for Approximate Bayesian Inference," in *Proc. Conf. Uncertainty in Artificial Intelligence*, ser. UAI'01, San Francisco, CA, USA, 2001, pp. 362–369.

[10] J. Céspedes, P. M. Olmos, M. Sanchez-Fernandez, and F. Perez-Cruz, "Expectation Propagation Detection for High-Order High-Dimensional MIMO Systems," *IEEE Trans. Commun.*, vol. 62, no. 8, pp. 2840–2849, Aug. 2014.

[11] Y. Zhang, Z. Wu, C. Li, Z. Zhang, X. You, and C. Zhang, "Expectation Propagation Detection with Neumann-Series Approximation for Massive MIMO," in *Proc. IEEE Int. Workshop Signal Process. Syst. (SiPS)*, Oct. 2018, pp. 59–64.

[12] G. Yao, G. Yang, J. Hu, and C. Fei, "A Low Complexity Expectation Propagation Detection for Massive MIMO System," in *Proc. IEEE Global Comm. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–6.

[13] W. Tang, H. Prabhu, L. Liu, V. Öwall, and Z. Zhang, "A 1.8 Gb/s 70.6 pJ/b 128× 16 link-adaptive near-optimal massive MIMO detector in 28nm UTBB-FDSOI," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2018, pp. 224–226.

[14] F. Wang, C. Zhang, J. Yang, X. Liang, X. You, and S. Xu, "Efficient matrix inversion architecture for linear detection in massive MIMO systems," in *Proc. IEEE Int. Conf. Digit. Signal Process (DSP)*, 2015, pp. 248–252.

[15] A. Habegger, A. Stahel, J. Goette, and M. Jacomet, "An efficient hardware implementation for a reciprocal unit," in *IEEE Int. Symp. Electron. Design, Test & Appl.*, 2010, pp. 183–187.

[16] G. Peng, L. Liu, S. Zhou, S. Yin, and S. Wei, "A 1.58 Gbps/W 0.40 Gbps/mm2 ASIC implementation of MMSE detection for 128 × 8 64-QAM Massive MIMO in 65 nm CMOS," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 65, no. 5, pp. 1717–1730, 2017.

[17] Y.-T. Chen, W.-C. Sun, C.-C. Cheng, T.-L. Tsai, Y.-L. Ueng, and C.-H. Yang, "An integrated message-passing detector and decoder for Polar-coded massive MU-MIMO systems," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 66, no. 3, pp. 1205–1218, 2018.