An Efficient Conjugate Residual Detector for Massive MIMO Systems

Yufeng Yang[1,2], Ye Xue[1,2], Xiaohu You[2], and Chuan Zhang[1,2,*]
[1]Lab of Efficient Architectures for Digital-communication and Signal-processing (LEADS)
[2]National Mobile Communications Research Laboratory, Southeast University, Nanjing, China
Email: {yfyang, yexue, xhyu, chzhang}@seu.edu.cn

***Abstract*—In nowadays wireless communication systems, massive multiple-input multiple-output (MIMO) technique brings better energy efficiency and coverage but higher computational complexity than small-scale MIMO. For linear detection such as minimum mean square error (MMSE), prohibitive complexity lies in solving large-scale linear equations. For a better trade-off between BER performance and computational complexity, iterative linear methods like conjugate gradient (CG) have been applied for massive MIMO detection. By leaving out a matrix-vector product of CG, conjugate residual (CR) further achieves lower computational complexity with similar BER performance compared to CG. Since the BER performance can be improved by utilizing pre-condition with incomplete Cholesky (IC) factorization, pre-conditioned conjugate residual (PCR) is proposed. Simulation results indicate that PCR method achieves better performance than both CR and CG methods. It has 1 dB performance improvement than CG at BER $= 5 \times 10^{-3}$. Analysis shows that CR achieves 20% computational complexity reduction compared with CG when antenna configuration is 128×60. With the same configuration, PCR reduces complexity by 66% while achieves similar BER performance compared with the detector with Cholesky decomposition. Finally, the corresponding VLSI architecture is proposed in detail.**

***Keywords*—Massive MIMO detection, conjugate gradient, conjugate residual, pre-condition, VLSI.**

I. Introduction

Massive multiple-input multiple-output (MIMO) is a key technique in for the next generation wireless systems [1] and has been incorporated into wireless broadband standards like the 3rd generation partnership project (3GPP) long term evolution (LTE) and IEEE 802.11n [2]. By equipping hundreds of antennas at transmitters and serving relatively a small number of users [3], massive MIMO provides significant improvement in spectral efficiency, interference reduction, transmit-power efficiency, and link reliability [4] compared with the conventional small-scale MIMO, where point-to-point MIMO links are mainly focused.

Because of the large antenna number at base stations (BSs) or user side, computational complexity becomes unaffordable in massive MIMO detection. Among existing detections, zero forcing (ZF) is a basic way, which neglects the effect of noise [5]. However, its performance is not satisfactory. Though linear scheme like minimum mean square error (MMSE) [6] tries to make a trade-off between performance and complexity, its computation complexity still increases drastically as the number of antennas grows. For a massive MIMO system with $N \times M$ dimensional channel matrix $\mathbf{H}$, computational complexity of MMSE inversion is $\mathcal{O}(M^3)$, which makes it costly in real application [7–10]. Thus, iterative linear solvers are proposed for further complexity reduction [11–13]. For instance, methods like Gauss-Seidel [11] and conjugate gradient (CG) are adopted, among which CG effectively optimizes the MMSE scheme with reduced computational complexity of $\mathcal{O}(M^2)$ with negligible performance loss.

It is believed that algorithms for massive MIMO detection can be further improved for either lower complexity or better performance. Considering the computation process of CG, some steps can be eliminated in specific way. CG involves three matrix-vector multiplications. For lower complexity, by leaving out a matrix-vector multiplication of CG, conjugate residual (CR) is proposed in this paper, which is also capable of solving Hermitian problems in general. However, straightforward application of CR in massive MIMO detection does not offer better performance than CG in some cases. Therefore, an improved CR is proposed by introducing pre-conditioner, named PCR. The pre-condition algorithm called incomplete Cholesky (IC) factorization is introduced for CR. Numerical results of the aforementioned three methods have demonstrated the advantages of the proposed CR and PCR methods. Since the MMSE filtering matrix is diagonal-dominant for uplink massive MIMO systems, the iteration number is chosen to be 2, 3, or 4, respectively. Finally, the hardware architecture of PCR method is also given.

The remainder of the paper is organized as follows. Section II goes over the system model and MMSE detection of massive MIMO detection. Section III presents the proposed CR-based algorithm and the pre-conditioned version of it. Performance comparisons are given in Section IV. Section V elaborates the computational complexity and hardware architecture. Finally, Section VI concludes the entire paper.

Notation: In this paper, the lowercase and upper bold face letters stand for column vector and matrix, respectively. The operations $(.)^T$ and $(.)^H$ denote transpose and conjugate transpose, respectively. The entry in the i-th row and j-th column of $\mathbf{A}$ is $\mathbf{A}_{(i,j)}$. The vector $\boldsymbol{\alpha}$ in the k-th iteration is $\boldsymbol{\alpha}_k$. Computational complexity is denoted in terms of complexed-valued multiplication number of the algorithm.

II. Preliminaries

Consider an uplink of an orthogonal frequency-division multiplexing (OFDM)-based massive MIMO system with N antennas at the base station, which simultaneously serves M single antenna users. Here, N is always much bigger than M ($N >> M$). The transmitted signal vector and received vector

978–1–5386–0446–5/17/$31.00

are denoted by $\mathbf{s} = [s_1, s_2, ..., s_M]^T$ and $\mathbf{y} = [y_1, y_2, ..., y_N]^T$, respectively, where $\mathbf{s} \in \mathbb{C}^M$, $\mathbf{y} \in \mathbb{C}^N$. Then the system model is described as

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

where $\mathbf{H}$ is an $N \times M$ uplink channel matrix, $\mathbf{n}$ is the vector representing Additive White Gaussian Noise (AWGN) with zero-mean and variance σ^2. According to MMSE scheme, At the base station side, the estimate $\hat{\mathbf{s}}$ of the transmitted symbol vector is

$$\hat{\mathbf{s}} = (\mathbf{H}^H\mathbf{H} + \sigma^2\mathbf{I}_M)^{-1}\mathbf{H}^H\mathbf{y} = \mathbf{A}^{-1}\tilde{\mathbf{y}}, \tag{2}$$

where the matrix $\mathbf{I}$ means identity matrix with dimension M, and the MMSE filtering matrix $\mathbf{A}$ is defined base on *Gram* matrix $\mathbf{G}$:

$$\mathbf{A} = \mathbf{G} + \sigma^2\mathbf{I}_M, \tag{3}$$

where $\mathbf{G} = \mathbf{H}^H\mathbf{H}$.

Correspondingly, output of matched filter $\tilde{\mathbf{y}}$ is

$$\tilde{\mathbf{y}} = \mathbf{H}^H\mathbf{y}. \tag{4}$$

Nevertheless, computational complexity of exact matrix inversion $\mathbf{A}^{-1}$ is $\mathcal{O}(M^3)$. Methods such as Cholesky decomposition based method are not suitable for adoption in massive MIMO detection when the scale of it increases.

III. Proposed Conjugate Residual Method

In this section, CG method for MMSE linear detection is firstly introduced. Then a new low-complexity method with the prototype of CG is proposed, named CR method, which can be applied in solving Hermitian problems. By adopting IC factorization, an improvement algorithm of CR is proposed as PCR, which is capable of better performance.

A. Conjugate Gradient Method

As mentioned in Section II. MMSE linear detection can be denoted by equation $\hat{\mathbf{s}} = \mathbf{A}^{-1}\tilde{\mathbf{y}}$, which can be seen as the typical linear problem. CG is a method that can be used in such problem [14], and it has its advantage by lowering the computational complexity.

Additionally, for massive MIMO uplink, channel matrix $\mathbf{H}$ is asymptotically orthogonal, which means that the MMSE filtering matrix is diagonally dominant. Thus, the channel matrix can be viewed as sparse and has large scale [15]. CG is a method for solving symmetric positive definite (SPD) matrices while filtering matrix $\mathbf{A}$ can be proved to be SPD. In conclusion, CG is a feasible method to reduce computational complexity from $\mathcal{O}(M^3)$ to $\mathcal{O}(M^2)$ by substituting the inverse with iteration method while achieves acceptable BER performance.

B. Proposed CR-based Detection

Consider CG invloves too much complex multiplications, computation process can be simplified by leaving out a matrix-vector product in every iteration of CG, CR is proposed as a new iterative method [16] for massive MIMO detection. As is mentioned above, the solution to Eq. (2) can be computed by solving the following optimization problem,

$$\hat{\mathbf{s}} = \underset{\hat{\mathbf{s}} \in \mathbb{C}^{\mathbf{M}}}{\arg\min} \|\tilde{\mathbf{y}} - \mathbf{A}\hat{\mathbf{s}}\|, \tag{5}$$

where $\mathbf{A} = \mathbf{H}^H\mathbf{H} + \sigma^2\mathbf{I}_M$ is the regularized uplink Gram matrix. Similar to CG, CR computes $\hat{\mathbf{s}}$ iteratively rather than computes the exact inverse of matrix $\mathbf{A}$, which is the advantage compared to Cholesky Inversion. Also, computational complexity is effectively reduced compared with CG. The solution to Eq. (3) can be computed (or approximated) by using CR algorithm, and computation process of CR is shown in *Algorithm* 1.

Algorithm 1 CR for MMSE Detection

Input: $\mathbf{A}$ and $\tilde{\mathbf{y}}$

$\mathbf{b} = \tilde{\mathbf{y}}$, $\mathbf{v}_0 = \mathbf{0}$, $\mathbf{r}_0 = \mathbf{b}$, $\mathbf{p}_0 = \mathbf{r}_0$
$\mathbf{e}_0 = \mathbf{A}\mathbf{p}_0$, $\mathbf{m}_0 = \mathbf{A}\mathbf{r}_0$
for $k = 1, \ldots, K$ **do**
  $\boldsymbol{\alpha}_k = \mathbf{r}_{k-1}^H\mathbf{m}_{k-1}/\|\mathbf{e}_{k-1}\|^2$
  $\mathbf{v}_k = \mathbf{v}_{k-1} + \boldsymbol{\alpha}_k\mathbf{p}_{k-1}$
  $\mathbf{r}_k = \mathbf{r}_{k-1} - \boldsymbol{\alpha}_k\mathbf{e}_{k-1}$
  $\mathbf{m}_k = \mathbf{A}\mathbf{r}_k$
  $\boldsymbol{\beta}_k = \mathbf{r}_k^H\mathbf{m}_k/\mathbf{r}_{k-1}^H\mathbf{m}_{k-1}$
  $\mathbf{p}_k = \mathbf{r}_k + \boldsymbol{\beta}_k\mathbf{p}_{k-1}$
  $\mathbf{e}_k = \mathbf{m}_k + \boldsymbol{\beta}_k\mathbf{e}_{k-1}$
end for

Output: $\hat{\mathbf{s}} = \mathbf{v}_k$

Algorithm 1 summarizes the CR-based data detection for massive MIMO system. From the derivation of CR method, $\mathbf{A}\mathbf{p}_j$ is replaced by $\mathbf{e}_j$ in each iteration, which can leave out the matrix-vector product in the algorithm of CG with the objective of lower complexity.

Besides, CR-based method can be applied when the system is Hermitian, while CG is a method for solving SPD problems. In contrast, CR is more suitable for massive MIMO systems.

C. Comparison between CG and CR

Comparison between CR and CG is mainly analyzed in three parts in this section.

1) Computation Method: Consider the initial estimate of $\hat{\mathbf{s}}$ to the MMSE linear detection is $\hat{\mathbf{s}}_0$, and $\mathbf{r}_k = \tilde{\mathbf{y}} - \mathbf{A}\hat{\mathbf{s}}_k$ is the residual vector for the approximation $\hat{\mathbf{s}}_k$ within the k-th Krylov subspace. The exact solution of the problem is denoted by $\hat{\mathbf{s}}^*$. CG is a method to solve MMSE linear problem by minimizing the function $\|\hat{\mathbf{s}}^* - \hat{\mathbf{s}}_k\|_A^2 \equiv (\hat{\mathbf{s}}^* - \hat{\mathbf{s}}_k)^T\mathbf{A}(\hat{\mathbf{s}}^* - \hat{\mathbf{s}}_k)$, known as the energy norm of error. CR is a method to minimize $\|\mathbf{r}_k\| = \|\tilde{\mathbf{y}} - \mathbf{A}\hat{\mathbf{s}}_k\|$. From the differences between CR and CG it could be seen that CR is proposed to deal with the residual energy, which explains the name of the method in the physical way.

2) Performance: CR is a algorithm from the prototype of CG. Thus, two algorithms have similar performances. However, as the iteration time increases, performance gap between two methods reduces gradually.

3) Complexity: Computational complexity reduction is the main characteristic of the proposed CR method. With the reduction of one matrix-vector product in each iteration, complexity is effectively lowered. To elaborate the improvement, CR need nearly the same complexity to CG when iteration time is 4 while that of CG is 3. Despite the latency, CR can achieve better performance with the same complexity.

D. Pre-Conditioned CR Method

Better performance is needed in practical application, while CR may not competitable in sense of performance. Thus an improvement of CR is necessary, pre-condition is a good way to improve the performance of specific algorithms, then incomplete Cholesky (IC) factorization is used to pre-condition the CR system matrix [17]. That is, matrix $\mathbf{A}$ can be pre-conditioned by IC method like [12]. To pre-condition the system matrix, a pre-conditioner matrix $\mathbf{M}$ can be utilized, then the linear equation can be rewritten as,

$$\mathbf{M}^{-1}\mathbf{A}\hat{\mathbf{s}} = \mathbf{M}^{-1}\tilde{\mathbf{y}}. \tag{6}$$

In SPD system, pre-conditioner can be written as $\mathbf{M} = \mathbf{L}\mathbf{L}^T$. In pre-condition process, pre-conditioner $\mathbf{L}\mathbf{L}^H$ is calculated efficiently to approach the original matrix. The function of pre-condition is to make the matrix optimal to be disposed.

$$(\mathbf{L}\mathbf{L}^H)^{-1}\mathbf{A} \approx \mathbf{I}, \tag{7}$$

where $\mathbf{I}$ is an identity matrix.

To determine the set of the pre-conditioned matrix, a specific way should be adopted. Consider the base-station-to-user ratio varies, this ratio is taken into consideration. Also, a constant parameter ν is determined by the requirements of performances. Thus the threshold value can be defined as,

$$\delta = \nu(1 - M/N)\mathbf{A}_{(i,i)}. \tag{8}$$

Take the threshold value δ into calculation, then the computation process of IC factorization is shown as *Algorithm* 2.

Algorithm 2 Incomplete Cholesky Factorization

Input: $\mathbf{A}$, δ, and $\frac{M}{N}$
for $j = 1:1:M$ **do**
  for $i = j:1:M$ **do**
    if $\mathbf{A}_{(i,j)} \leq \delta$ **then**
      $\mathbf{L}_{(i,j)} = 0$
    else
      $\Sigma = 0$
      for $k = 1:1:j-1$ **do**
        $\Sigma = \Sigma + \mathbf{L}_{(i,k)}\mathbf{L}_{(j,k)}/\mathbf{L}_{(k,k)}$
      end for
      $\mathbf{L}_{(i,j)} = \mathbf{A}_{(i,j)} - \Sigma$
    end if
  end for
end for
Output: $\mathbf{L}$

After the calculation of $\mathbf{L}$ in *Algorithm* 2, detection result in traditional CR method can be easily calculated with pre-condition. Because matrix $\mathbf{M}$ serves as a pre-conditioner in computation process, then PCR algorithm is as *Algorithm* 3.

Algorithm 3 Proposed PCR Method

Input: $\mathbf{A}$, $\mathbf{L}$ and $\mathbf{y}$
$\mathbf{b} = \tilde{\mathbf{y}}$, $\mathbf{v}_0 = \mathbf{0}$, $\mathbf{r}_0 = \mathbf{L}^{-1}\mathbf{b}$, $\mathbf{p}_0 = (\mathbf{L}\mathbf{L}^H)^{-1}\mathbf{b}$
$\mathbf{m}_0 = \mathbf{A}\mathbf{r}_0$, $\mathbf{e}_0 = \mathbf{A}\mathbf{p}_0$
for $k = 1, \ldots, K$ **do**
  $\boldsymbol{\alpha}_k = \mathbf{r}_{k-1}^H\mathbf{m}_{k-1}/\|\mathbf{e}_{k-1}\|^2$
  $\mathbf{v}_k = \mathbf{v}_{k-1} + \boldsymbol{\alpha}_k\mathbf{p}_{k-1}$
  $\mathbf{r}_k = \mathbf{r}_{k-1} - \boldsymbol{\alpha}_k\mathbf{L}^{-1}\mathbf{e}_{k-1}$
  $\mathbf{m}_k = \mathbf{A}\mathbf{r}_k$
  $\beta_k = \mathbf{r}_k^H\mathbf{m}_k/\mathbf{r}_{k-1}^H\mathbf{m}_{k-1}$
  $\mathbf{p}_k = \mathbf{L}^{-H}\mathbf{r}_k + \beta_k\mathbf{p}_{k-1}$
  $\mathbf{e}_k = \mathbf{A}\mathbf{p}_k$
end for
Output: $\hat{\mathbf{s}} = \mathbf{v}_k$

E. Proof of Convergence

The estimate $\hat{\mathbf{s}}_k$ can be proved to support the convergence of the algorithm [18]. For CR on a SPD system,

$$\|\hat{\mathbf{s}}_k\|^2 - \|\hat{\mathbf{s}}_{k-1}\|^2 = 2\alpha_k\hat{\mathbf{s}}_{k-1}^T\mathbf{p}_{k-1} + \mathbf{p}_{k-1}^T\mathbf{p}_{k-1} \geq 0. \tag{9}$$

Therefore,

$$\|\hat{\mathbf{s}}_k\| \geq \|\hat{\mathbf{s}}_{k-1}\|. \tag{10}$$

Then, final solution can be expressed as $\hat{\mathbf{s}}_l = \hat{\mathbf{s}}^*$,

$$\begin{aligned}\mathbf{s}_l &= \mathbf{s}_{l-1} + \alpha_{l-1}\mathbf{p}_{l-1}\\ &= \cdots\\ &= \hat{\mathbf{s}}_k + \alpha_{k+1}\mathbf{p}_k + \cdots + \alpha_{l-1}\mathbf{p}_{l-1}\\ &= \hat{\mathbf{s}}_{k-1} + \alpha_k\mathbf{p}_{k-1} + \alpha_{k+1}\mathbf{p}_k + \cdots + \alpha_{l-1}\mathbf{p}_{l-1}.\end{aligned} \tag{11}$$

From the conclusion above, it can deduced that

$$\begin{aligned}&\|\hat{\mathbf{s}}_l - \hat{\mathbf{s}}_{k-1}\|^2 - \|\hat{\mathbf{s}}_l - \hat{\mathbf{s}}_k\|^2\\ &= (\hat{\mathbf{s}}_l - \hat{\mathbf{s}}_{k-1})^T(\hat{\mathbf{s}}_l - \hat{\mathbf{s}}_{k-1}) - (\hat{\mathbf{s}}_l - \hat{\mathbf{s}}_k)^T(\hat{\mathbf{s}}_l - \hat{\mathbf{s}}_k)\\ &= 2\alpha_k\mathbf{p}_{k-1}^T(\alpha_{k+1}\mathbf{p}_k + \cdots + \alpha_{l-1}\mathbf{p}_{l-1})\\ &+ \alpha_k^2\mathbf{p}_{K-1}^T\mathbf{p}_{k-1} \geq 0.\end{aligned} \tag{12}$$

While for the MMSE linear detection problem, the linear equation $\tilde{\mathbf{y}} = \mathbf{A}\hat{\mathbf{s}}$ is to be solved, thus

$$\begin{aligned}&\|\hat{\mathbf{s}}_l - \hat{\mathbf{s}}_{k-1}\|_{\mathbf{A}}^2 - \|\hat{\mathbf{s}}_l - \hat{\mathbf{s}}_k\|_{\mathbf{A}}^2\\ &= (\hat{\mathbf{s}}_l - \hat{\mathbf{s}}_{k-1})^T\mathbf{A}(\hat{\mathbf{s}}_l - \hat{\mathbf{s}}_{k-1}) - (\hat{\mathbf{s}}_l - \hat{\mathbf{s}}_k)^T\mathbf{A}(\hat{\mathbf{s}}_l - \hat{\mathbf{s}}_k)\\ &= 2\alpha_k\mathbf{p}_{k-1}^T\mathbf{A}(\alpha_{k+1}\mathbf{p}_k + \cdots + \alpha_{l-1}\mathbf{p}_{l-1})\\ &+ \alpha_k^2\mathbf{p}_{k-1}^T\mathbf{A}\mathbf{p}_{k-1}\\ &= 2\alpha_k\mathbf{q}_{k-1}^T(\alpha_{k+1}\mathbf{p}_k + \cdots + \alpha_{l-1}\mathbf{p}_{l-1})\\ &+ \alpha_k^2\mathbf{q}_{k-1}^T\mathbf{p}_{k-1} > 0.\end{aligned} \tag{13}$$

The derivation above indicates that the error in energy norm is strictly decreasing. Hence, the method is valid for massive MIMO detection.

F. Complexity Analysis

CR leaves out a matrix-vector product in calculation of $\boldsymbol{\alpha}$ in each iteration of CG, thus complexity of CR method is reduced. To be clear, M^2 complex-value multiplications are reduced compared with CG and M multiplications are added in each iteration due to new vector storage.

After IC factorization, performance complexity trade-off makes complexity of PCR higher, so new complexity is added to CR. Meanwhile, added complexity is complex, thus complexity of PCR method is analyzed in detail and each step is considered as follow:

1) Initialization: initializing $\mathbf{r}$, $\mathbf{p}$ and $\mathbf{e}$ needs complexity. Take pre-condition into consideration, Q is the number of zeros in the lower triangle matrix, thus number of complex multiplication is: $3M^2 - M - 2Q$.

2) Calculating $\mathbf{v}$ *and* $\mathbf{r}$*:* the complexity needed in calculating $\boldsymbol{\alpha}$ is two vector-vector products and a matrix-vector product, while calculating $\mathbf{v}$ needs one more vector-vector product and calculating $\mathbf{r}$ needs one product. So the total number of complex multiplication is: $5M$.

3) Calculating $\mathbf{p}$ *and* $\mathbf{e}$*:* the complexity needed is $2M^2 - 2Q$ for calculating $\mathbf{m}_k$ and $\boldsymbol{\beta}$, then calculating the $\mathbf{p}$ and $\mathbf{e}$ needs $2M$. Thus number of complex multiplication is: $2M^2 + 2M - 2Q$.

4) Pre-condition: the pre-condition complexity is from the IC factorization, whose number of complex multiplication is: $((2M^2 - 2Q)^{3/2} - (M^2 - 2Q)^{1/2})/6$.

IV. Performance Comparison

CR is derivated from CG and PCR is an improved CR algorithm, thus performance of three methods and the Cholesky exact MMSE detection is compared. Consider the massive MIMO systems with antenna configurations of $N \times M = 128 \times 32$, $N \times M = 128 \times 16$ and $N \times M = 128 \times 8$. The modulation scheme is 64-QAM. Here $\nu = 1/9$. k is the iteration number. The performance is shown in Fig.s 1, 2, and 3.

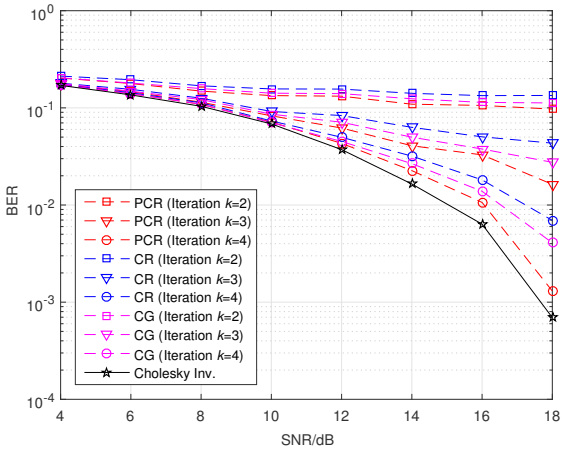

Fig. 1. BER performance comparison with $N \times M = 128 \times 32$.

According to Fig.s 1, 2, and 3, as ratio $N \times M$ increases, difference between CR and CG is decreasing while that between PCR and CG is increasing. Essence of algorithms are different, and CR is to minimize the residual energy of the detection while CG is to minimize the estimation of the result, high iteration count contribute to the performance of CR. Better initial condition improves the performance of PCR.

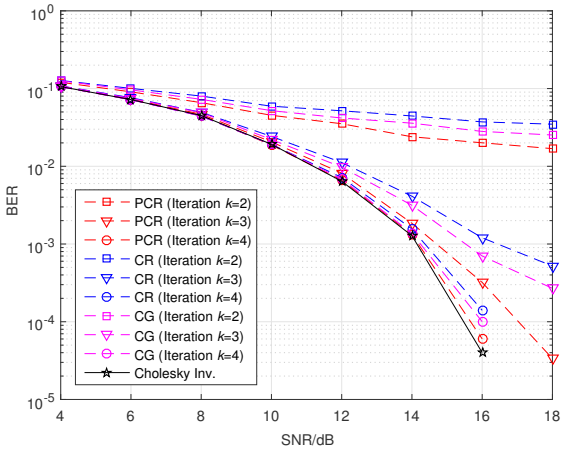

Fig. 2. BER performance comparison with $N \times M = 128 \times 16$.

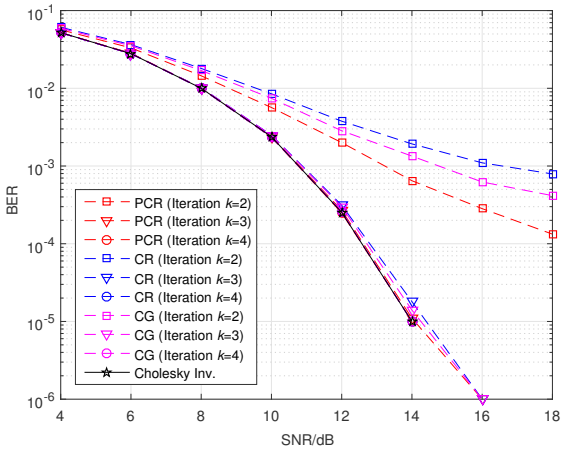

Fig. 3. BER performance comparison with $N \times M = 128 \times 8$.

From three figures, performance of CR can be found that CR has similar performance with CG because the prototype of CR is CG. While as the iteration time increases, the curve of CR converges quickly. For iteration time $k \geq 3$, CR method has almost the same performance with CG. PCR method achieves better performance than both CR and CG, which corresponds to the objective of application of pre-condition.

Low-complexity algorithm CR achieves similar performance to CG. In Fig. 1, it has 0.4dB performance drawback at BER$= 10^{-2}$ when iteration time $k = 4$. However, as ratio M/N decreases, difference between CR and CG drops, which can be seen in Fig. 3 that when iteration time $k \geq 3$, CR achieves nearly the same performance with CG. Trade-off by lowering complexity is not significant in these cases.

PCR has better performance than both CG and CR in each configuration. At BER$= 10^{-3}$, the SNR gap is 0.6dB

in comparison with CG in Fig. 2, while at BER$= 5 \times 10^{-3}$, SNR gap is 1dB in comparison with CG in Fig. 1. While in Fig. 3, PCR have its SNR advantage over CG by 2dB at BER$= 1 \times 10^{-3}$. It is obvious that under different ratio M/N, performance of PCR is better than CG.

V. Hardware Architecture

In this section, computational complexities are compared to elaborate both CR and PCR. VLSI architecture is also introduced with processing steps.

A. Computational Complexity

In this section, computational complexity is analyzed in terms of numbers of complex multiplications of the algorithm. Computational complexity of CR and PCR are elaborated in contrast with CG. As is mentioned, CR has its advantage by leaving out a matrix-vector product and adopt a new vector as a substitution. After this transform, the algorithm has a change in essence from minimizing the result to minimizing the residual.

To improve BER performance, PCR achieves better performance than both CR and CG. However, performance and complexity trade-off makes the computational complexity increases at the base of CR due to the pre-condition. The exact value of computational complexity is summarized in *Table* I to show the difference between three methods.

TABLE I. Required Number of Complex Multiplications

	Symbol $\hat{\mathbf{s}}$
Cholesky Inversion	$\frac{5}{6}M^3+\frac{3}{4}M^2+\frac{4}{3}M$
CG	$k(2M^2+6M)$
CR	$k(M^2+7M)+2M^2$
PCR *	$\frac{(M^2-2Q)^{3/2}-(M^2-2Q)^{1/2}}{6}+$ $k(2M^2+7M-2Q)+3M^2-M-2Q$

* calculates the complexity of PCR method and the variable Q means the number of zero in the lower triangle matrix $\mathbf{L}$.

To make the comparison more explicit, complexity comparisons of three methods are elaborated in Section V-A1 and Section V-A2. Consider number of antennas at the base station N is 128, SNR is set to be 20dB, constant parameter ν is $1/9$ in terms of PCR.

1) Complexity of CR: With the omission of specific product, CR can reduce the complexity compared with CG. As is shown in Fig. 4, complexity of CR is reduced by nearly 20% compared with CG when number of antennas at the user side M is 60. Meanwhile CR achieves almost the same complexity when iteration time $k = 4$ with CG of iteration time $k = 3$.

2) Complexity of PCR: With the affiliation of pre-condition, complexity of PCR increases at the base of CR. To achieve relatively higher performance, necessary trade-off is made. Complexity of PCR is higher than CG while can lower the complexity of Cholesky inversion by nearly 66% when $M = 60$, as is shown in Fig. 5.

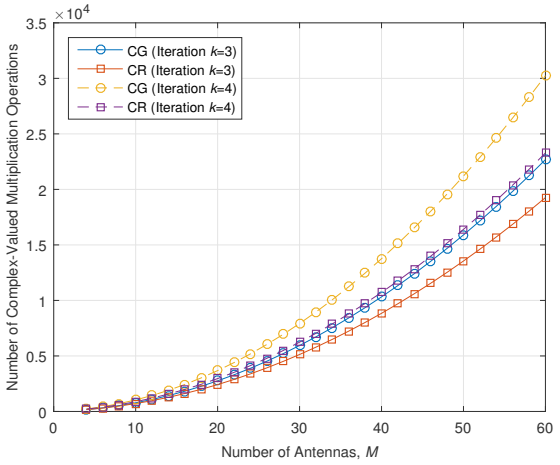

Fig. 4. Complexity comparison between CR and CG.

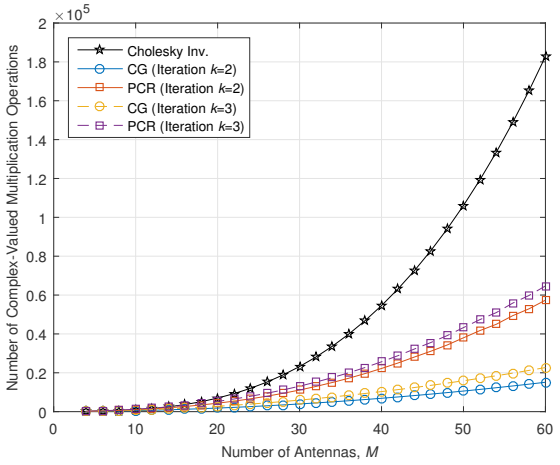

Fig. 5. Complexity comparison between PCR and CG.

B. Hardware Architecture

Proposed hardware architecture shown in Fig. 6 includes three units: 1) preprocessing unit, 2) proposed PCR method unit, and 3) output unit. Preprocessing unit computes $\tilde{\mathbf{y}}$ and $\mathbf{A}$, and $\tilde{\mathbf{y}}$ is denoted by $\mathbf{y}^E$ in the architecture. Proposed PCR method unit and output unit compute and output $\hat{\mathbf{s}}$ iteratively, while $\hat{\mathbf{s}}$ is denoted by $\mathbf{s}^E$ in the architecture.

1) Preprocessing Unit: In preprocessing unit, output of matched filter $\tilde{\mathbf{y}}$ and channel matrix are processed to give *Gram* matrix and pre-conditioner $\mathbf{M}$. Consider $\mathbf{A}$ is Hermitian, $M \times M$ lower triangular systolic array is adopted to compute it. Each processing element (PE) performs a multiply-accumulate (MAC) operation with same inputs.

2) Proposed PCR Method Unit: In this unit, module INV means matrix inversion. Two inputs of module DIV are dividend on the left and divisor on the top. Mod module calculates the modulus of input. Module with $\mathbf{r}$, $\mathbf{m}$, $\mathbf{p}$, $\mathbf{e}$ store the corresponding value in the k-th iteration and the downside input means initialization. With the output of preprocessing unit, each iteration of PCR method unit has 6 phases.

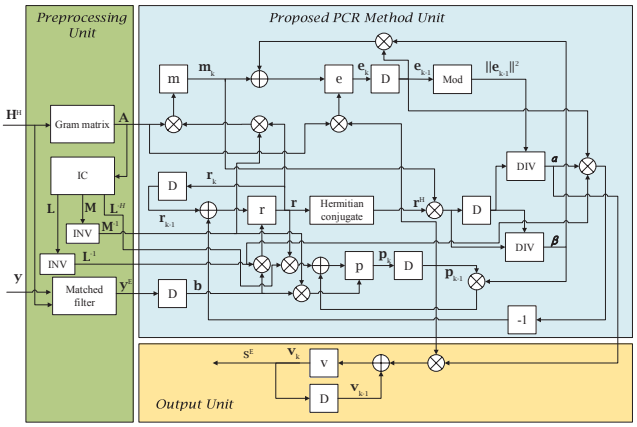

Fig. 6. Hardware architecture of PCR method.

a) In the first phase: With initialization, $\boldsymbol{\alpha}$ is calculated firstly after DIV module by output of multiplier of $\mathbf{r}^H$ and $\mathbf{m}$, along with $\|\mathbf{e}\|^2$, which is the output of Mod module .

b) In the second phase: $\mathbf{r}$ can update itself with the value of $\boldsymbol{\alpha}$ and $\mathbf{e}$ by one adder.

c) In the third phase: $\mathbf{m}$ is updated through two multipliers with $\mathbf{A}$ and $\mathbf{M}^{-1}$ and the updated $\mathbf{r}$.

d) In the fourth phase: Intermediate variable $\boldsymbol{\beta}$ is calculated by DIV module with inputs from the current iteration and the last iteration.

e) In the fifth phase: Vector $\mathbf{p}$ is updated by adder with inputs from $\mathbf{r}$, and multiplication of $\boldsymbol{\beta}$ and $\mathbf{p}$ from last iteration.

f) In the sixth phase: $\mathbf{e}$ is calculated for the next iteration by one adder and one multiplier.

3) Output Unit: In this unit, estimate $\hat{\mathbf{s}}$ are calculated by $\mathbf{v}$ iteratively, $\mathbf{v}$ updates itself with the output of PCR method unit $\mathbf{p}$ and $\boldsymbol{\alpha}$.

With the result in Section IV, output of proposed PCR method architecture is close to exact solution, thus iteration time does not need to be large, as mentioned in Section I.

VI. Conclusion

In this paper, CR is first introduced to improve the detection efficiency for massive MIMO systems. Both CG and CR are introduced and their relationship is shown as well. An improved version of CR is introduced called PCR to achieve better performance. Simulations are made to compare three methods and conventional detection based on exact matrix inversion. BER performances and computational complexities indicate that CR is efficient and suitable for low-complexity detection on massive MIMO while its improved version PCR can achieve higher BER performance in massive MIMO detection. Hardware architecture of PCR method is also proposed in the end. Future work will be directed towards the FPGA implementation of both CR and PCR detectors for our massive MIMO platform.

Acknowledgment

This work is partially supported by NSFC under grant 61501116, International Science & Technology Cooperation Program of China under grant 2014DFA11640, Huawei HIRP Flagship under grant YB201504, Jiangsu Provincial NSF under grant BK20140636, Intel Collaborative Research Institute for MNC, the Fundamental Research Funds for the Central Universities, State Key Laboratory of ASIC & System under grant 2016KF007, the Project Sponsored by the SRF for the Returned Overseas Chinese Scholars of State Education Ministry, and SRTP of Southeast University.

References

[1] E. Larsson *et al.*, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, 2014.
[2] J. Lee, J.-K. Han, and J. C. Zhang, "MIMO technologies in 3GPP LTE and LTE-advanced," *EURASIP Journal on Wireless Communications and Networking*, vol. 2009, no. 1, pp. 1–10, 2009.
[3] H. Huh, G. Caire, H. C. Papadopoulos, and S. A. Ramprashad, "Achieving "massive MIMO" spectral efficiency with a not-so-large number of antennas," *IEEE Trans. Wireless Commun.*, vol. 11, no. 9, pp. 3226–3239, 2012.
[4] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 742–758, 2014.
[5] Q. H. Spencer, A. L. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 461–471, 2004.
[6] E. G. Larsson, "MIMO detection methods: How they work [lecture notes]," *IEEE Signal Process. Mag.*, vol. 26, no. 3, 2009.
[7] A. Krishnamoorthy and D. Menon, "Matrix inversion using Cholesky decomposition," in *Proc. IEEE Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, 2013, pp. 70–72.
[8] 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 International Conference on Digital Signal Processing (DSP)*, July 2015, pp. 248–252.
[9] F. Wang, C. Zhang, X. Liang, Z. Wu, S. Xu, and X. You, "Efficient iterative soft detection based on polynomial approximation for massive MIMO," in *Proc. International Conference on Wireless Communications Signal Processing (WCSP)*, Oct 2015, pp. 1–5.
[10] X. Liang, C. Zhang, S. Xu, and X. You, "Coefficient adjustment matrix inversion approach and architecture for massive MIMO systems," in *Proc. IEEE International Conference on ASIC (ASICON)*, Nov 2015, pp. 1–4.
[11] Z. Wu, C. Zhang, Y. Xue, S. Xu, and X. You, "Efficient architecture for soft-output massive MIMO detection with Gauss-Seidel method," in *Proc. IEEE Circuits and Systems (ISCAS)*, 2016, pp. 1886–1889.
[12] Y. Xue, C. Zhang, S. Zhang, and X. You, "A fast-convergent preconditioned conjugate gradient detection for massive MIMO uplink," in *Proc. IEEE Digital Signal Processing (DSP)*, 2016, pp. 331–335.
[13] Y. Xue, C. Zhang, S. Zhang, Z. Wu, and X. You, "Steepest descent method based soft-output detection for massive MIMO uplink," in *Proc. IEEE International Workshop on Signal Processing Systems (SiPS)*, Oct 2016, pp. 273–278.
[14] B. Yin, M. Wu, J. R. Cavallaro, and C. Studer, "Conjugate gradient-based soft-output detection and precoding in massive MIMO systems," in *Proc. IEEE International Workshop on Green Communications, parallel with IEEE GLOBECOM*, 2014, pp. 3696–3701.
[15] J. A. Meijerink and H. A. van der Vorst, "An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix," *Mathematics of computation*, vol. 31, no. 137, pp. 148–162, 1977.
[16] Y. Saad, *Iterative methods for sparse linear systems*. SIAM, 2003.
[17] M. Benzi, "Preconditioning techniques for large linear systems: a survey," *Journal of computational Physics*, vol. 182, no. 2, pp. 418–477, 2002.
[18] D. C.-L. Fong and M. A. Saunders, "CG versus MINRES: An empirical comparison," *SQU Journal for Science*, vol. 17, no. 1, pp. 44–62, 2012.