PROCEEDINGS OF SCIENCE

A Novel Approximate-Exact Matrix Inversion Selection Method for Linear Data Detection in the Massive Multiple Input Multiple Output Uplink with Reconfigurable Implementation Results

Xiao Yang[1]
The Institute of Microelectronics, Tsinghua University, Beijing,100084, China
E-mail: davidyangcq@sina.com

Leibo Liu, Guiqiang Peng, Peng Zhang
The Institute of Microelectronics, Tsinghua University, Beijing,100084, China

Massive MIMO has become a key technology to the next wireless communication generation for its improvement in data rate, link reliability and power consumption; however, these benefits come at the cost of enormous data quantity, especially in the data detection of Massive MIMO uplink. In this paper, we propose a novel matrix inversion to balance the computational complexity and performance by providing selection of the 2-terms Neumann series approximation and the LDL decomposition matrix inversion methods according to different dimensions of the channel matrix. To reduce the hardware resource for two inversion algorithms and execute tasks efficiently, we consider the reconfigurable implementation for its flexibility and high-power efficiency. The implementation results are given by using our Reconfigurable Computing System for various antenna configurations. With this reconfigurable implementation, the throughput can achieve 93.8Mb/s, 130.4Mb/s, 82.2Mb/s and 107.1Mb/s for a 14×4 , 32×4 , 64×8 and 128×8 system respectively, which are chose to even better than few implementations for high dimension problems.

CENet2015
12-13 September 2015
Shanghai, China

[1] Speaker

PoS(CENet2015)021

1. Introduction

With the development of wireless communication technology, Multiple-Input-Multiple-Output (MIMO) has become a key technology in most modern wireless communication standards, such as 3GPP LTE [1-3] and IEEE 802.11n [4] because it offers improved link reliability, higher spectral efficiency and data rates compared to conventional single antenna systems [5]; however, since MIMO start approaching the bottleneck, a novel technology is required to meet the ever-growing demand for higher data rates [6-7].

Massive MIMO is an emerging technology which uses a large excess of service-antennas to provide more expansive channel space[8]. Huge improvements on throughput and link reliability can benefit from Massive MIMO [8, 9]; however, these benefits bring about some new problems in urgent need of attention and solution, especially the significantly increased computational complexity of data detection in the Massive MIMO uplink [8]. Since Minimum Mean Square Error (MMSE) detection is the most prominent low-complexity linear algorithm[10-11], we take it as the research object.

The remainder of the paper is organized as follows. Section 2 introduces the uplink signal model in Massive MIMO and its MMSE detection algorithm. Then we address the computational complexity and performance issue of two matrix inversion methods, Neumann series approximation and LDL decomposition, which represent approximate and exact matrix inversion respectively, and propose an approximate-exact matrix inversion selection method in section 3. Based on this selection method, Section 4 gives the reconfigurable implementation results of Neumann series approximation and LDL decomposition on our Reconfigurable Computing System, which carries out the calculations with received array configuration information from the host processor so that hardware resource is saved, high-dimension cases can be scaled to favorably without scale expansion of hardware and low-complexity in combination with optimal-performance is achieved.

2. Massive MIMO Uplink And MMSE Detection Algorithm

2.1 Massive MIMO Uplink Signal Model

A Massive MIMO uplink can be described as a system with N receiving antennas at the base station which receives signals from M transmitting antennas, which can be considered as M single antenna users. The transmitted data flow is divided into M sub-bit-streams, transmitted to receive antennas after each sub-bit-stream being encoded and mapped to constellation points. The receipt signal of each antennas is contained in the received symbol vector $y=[y_1,\ldots,y_N]^T$, which is given by

$$y=Hx+n \tag{2.1}$$

where x is a complex-valued vector described as $x=[x_1,\ldots,x_M]^T$ corresponding to M transmit symbols, H is the uplink channel matrix with dimension of $N\times M$, and vector n models the additive white Gaussian noise with variance $\sigma^2=N_0$. We furthermore assume that the transmission symbols satisfy $E\{|x_i|^2\}=E_s,\forall i$.

2.2 MMSE Soft-out Detection for Massive MIMO

The task of a data detector at the base station is to compute the estimates for the encoded sub-bit-streams given the vector y and the channel matrix H [12]. As to the MMSE detection, an estimate of the transmit vector x is computed as

$$\tilde{x}=(H^H\cdot H+N_0\cdot E_s^{-1}I)^{-1}\cdot H^H y \tag{2.2}$$

Formula (2.2) mainly consists of two parts, the $M\times M$ Gram matrix multiplication

$$G=H^H H \tag{2.3}$$

and the inversion of regularized matrix

$$A=G+N_0 E_s^{-1} I_M \quad . \tag{2.4}$$

3. Approximate-Exact Matrix Inversion Selection Method

Since the computational complexity is mainly caused by A^{-1} , a low-complexity inversion method without sacrificing the performance is a key point. In this section, we propose a novel method that enables the incorporation of an approximate, exacts matrix inversion, and presents the gist for selection of both methods according to various dimensions of channel matrix H by analyzing the computational complexity and BER performance of Neumann series approximation and LDL decomposition, which are typical cases of the approximation and exact inversion respectively.

3.1 Neumann Series Approximation and LDL Decomposition

According to Literatures, A^{-1} can be expressed by using the Neumann series as follows:

$$\tilde{A}_k^{-1}=\sum_{n=0}^{k-1}\left(-D^{-1}E\right)^n D^{-1} \tag{3.1}$$

in which D is the main diagonal matrix of A and E is the off-diagonal matrix[13, 14].

We take k=2 as the “2-terms Neumann series approximation” like

$$\tilde{A}_2^{-1}=D^{-1}-D^{-1}E\,D^{-1} \tag{3.2}$$

which only requires $O\left(M^2\right)$ operations in contrary to $O\left(M^3\right)$ of an exact algorithm.

In spite of the reduced complexity by virtue of this approximation method, it occurs inevitable error floor when N is not large enough as M.

LDL decomposition is an improved version of the classic Cholesky decomposition with less complexity. For this inversion method, A can be decomposed into $A=LDL^H$, in which L is a lower-triangular matrix with all main diagonal elements being 1. Make a middle matrix $V=LD$, namely $v_{ij}=l_{ij}\cdot d_j$, to simplify the data organization in memory; hence, there is

$$A=L\,V^H \quad . \tag{3.3}$$

L can be deduced column by column via expansion of Formula (3.3):

$$a_{ij}=\sum_{k=1}^{j} l_{ik}\,v_{jk}^{*}=\sum_{k-1}^{j-1} l_{ik}\left(j_{jk}\,d_k\right)^{*}+l_{ij}\,d_j\left(j\leqslant i\right) \tag{3.4}$$

Then we compute L^{-1} by $L\,L^{-1}=I$, namely,

$$l_{ij}=\sum_{k=j}^{i} l_{ik}\,l_{kj}^{-1} \quad . \tag{3.5}$$

Consequently, A^{-1} can be computed as:

$$A^{-1}=\left(L^{-1}\right)^H D^{-1} L^{-1} \quad . \tag{3.6}$$

3.2 Computational Complexity and BER Performance Comparison

Next we address the computational complexity and BER performance of both methods to make a comparison. Computational complexity consists of multiplication, addition and reciprocal operation. Since H is complex-valued, we convert the complex-valued operations to real-valued equivalents, concretely, 4 real multiplications and 2 additions for 1 non-conjugated complex multiplication, 2 real multiplications and 1 addition for 1 conjugated complex multiplication, and 2 real for 1 complex addition.

	Neumann Approximation (k=2)	LDL Decomposition
Multiplication	$4M(M-1)$	$2M(M-1)(M-2)/3$
Addition	$2M^2-M$	$2M^3+4M/3$
Reciprocal Operation	M	M

Table 1: Computational Complexity of 2-terms Neumann Series Approximation and LDL Decomposition

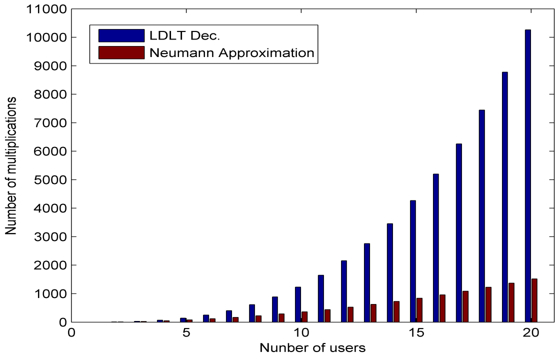


Figure 1: Complexity Comparison in Multiplication with Increasing Number of Users

Table 1 presents the complexity of 2-terms Neumann series approximation and LDL decomposition. The former has a lower complexity obviously; furthermore, with the channel matrix scaling up, the complexity gap becomes larger and larger as shown in Fig. 1.

Then we simulate an encoded system with the modulation scheme of 64 QAM to analyze the block error rate (BER) performance against the signal-to-noise ratio (SNR) for the both methods.

Fig. 2 characterizes the BER performance comparison between the two methods for M = 8 users. As to a certain M, the approximation inversion is approaching the performance of the exact inversion. In other words, the ratio of N and M is bigger, the performance of the approximation inversion will be better. Even so, we must pay attention to the error floor incurred by the approximation inversion in case N is close to M.

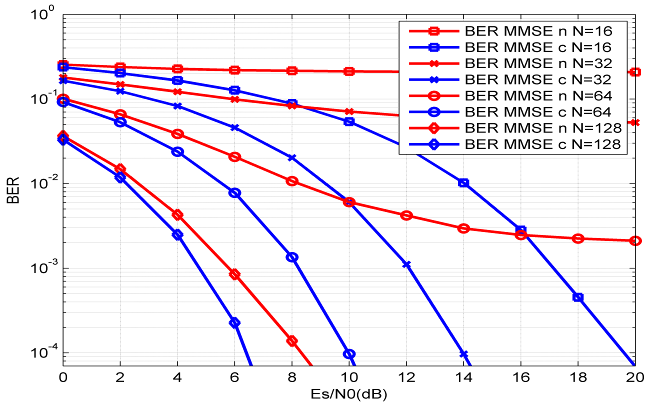
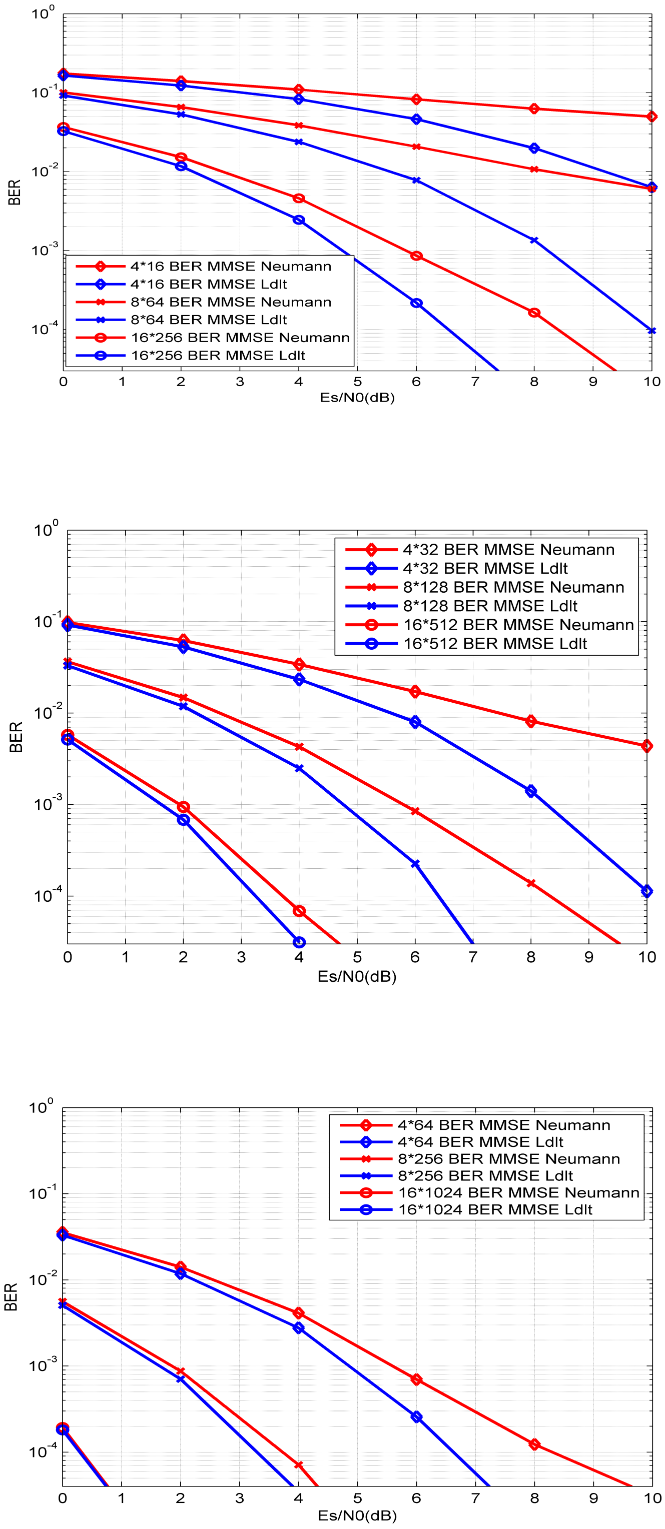


Figure 2: BER Performance Comparison for M=8 users (Transmitting Antennas)

3.3 A Novel Approximate-Exact Matrix Inversion Selection Method

Based on the investigation of the computational complexity and BER performance, as we can see, a selection of matrix inversion methods is necessary to balance the complexity and performance of different dimension cases. Next we depict the BER performance results of the approximate and exact methods, for a given N/M^2, and propose a method to estimate “a critical value” of N/M^2 to enable the selection.



(a) $N/M^2=1$

(b) $N/M^2=2$

(c) $N/M^2=4$

Figure 3: BER Performance Comparison for Different

The resulting BER performance is shown in Figs. 3(a), 3(b) and 3(c) for $N/M^2=1$, $N/M^2=2$, $N/M^2=4$ respectively. For a given N/M^2, the approximate inversion for a large-scale case significantly outperforms a small-scale case, and the performance varies with the increase of N/M^2, similarly to the literature [15].

We propose a method to measure the performance-gap between the approximate and exact inversion, i.e. the difference of SNRs for both methods required to achieve a certain BER. Fig. 4 characterizes this performance-gap against N/M^2 varying from 1 to 4 for M=4, M=8, and M=16 founding on 48 simulation points for each M. Every simulation point indicates the performance-gap for a certain dimension of channel matrix, and is measured at BER=, BER=, and BER= for M=4, M=8, and M=16 separately.

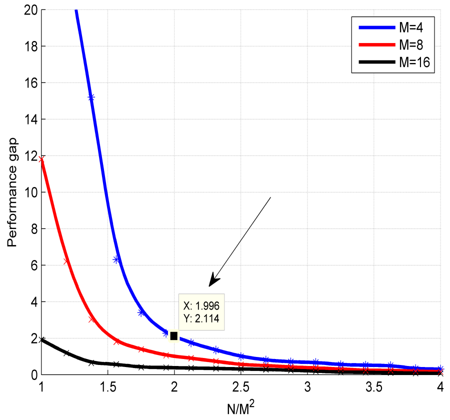


Figure 4: Performance-gap against N/M^2 and the Estimated "Critical Value" of N/M^2

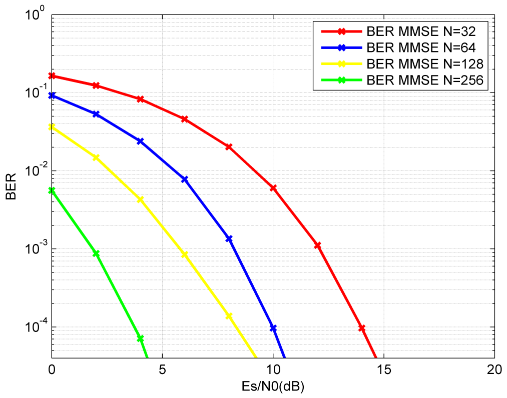


Figure 5: Performance of the Selection Method for M=8

It is obvious that if a point satisfies a low performance-gap for a smaller M, the bigger one will satisfy too.. Since there is no standard for this performance-gap to achieve a certain BER in MMSE, we give the $N/M^2=2$ at 2 dB performance-gap on M=4 as the estimated "critical value" for the selection of the approximate and exact inversion methods. Fig. 5 is the performance of MMSE detection by using this selection method for M=8, when $N/M^2\geqslant 2$ the LDL decomposition is utilized. Conversely, the approximate method is utilized.

4. Reconfigurable Implementation Results

In recent years, various reconfigurable architectures were proposed in Literatures [9-10]. Our reconfigurable system consists of 4 arrays of processing elements (PEs) and a host processor. The PE arrays handle parallel tasks by receiving configurable information from the host processor to achieve high throughput and low hardware consumption. In this case, as the matrix multiplication has the advantage of huge parallelism and two inversion methods are required to implemented, we consider the reconfigurable implementation and give the implementation results.

Operation	Antenna config.	This work (approximate-exact matrix inversion)			2-terms Neumann series approximation		
		freq. (MHz)	latency (cycles)	throughput (M matrix/s)	max freq.	latency	throughput
Gram matrix multiplication	16×4	250	21	11.90			
	32×4		37	6.76	302.29	46	6.57
	64×8		79	3.16			
	128×8		112	2.23	299.76	150	2
matrix inversion	16×4	250	64	3.91			
	32×4		46	5.43	301.57	52	5.8
	64×8		146	1.71			
	128×8		57	4.39	285.46	55	5.18

Table 2: Implementation Results by Using Our Reconfigurable Computing System

Table 2 shows the implementation results for different antenna configurations. As 2-terms Neumann series approximation is proposed in Literature [14], we put it as our comparison. In Table 1, we have considered two different cases of antenna configuration for $N/M^2=1$ and $N/M^2=2$, and give the latency required to deal with one matrix. All operations are carried out under a constant clock frequency of 250MHz based on our reconfigurable system. The throughput can be generally given by

$$Throuput=Frequency/Latency\ . \tag{4.1}$$

We can see that our implementation of Gram matrix multiplication achieves slightly greater throughput being 6.76 and 2.23 M matrixes/s compared with 6.57 and 2.00 M matrixes/s in Literature[14] for 16×4 and 32×4 , 64×8 and 128×8 that of matrix inversion for $N/M^2=2$ is close to Literature[14]. As there is rarely literature focused on the implementation of exact matrix inversion with such high dimension problems, no comparison can be given for $N/M^2=1$; however, it's obvious that the exact inversion consumes more clock cycles than the approximate inversion, and Gram matrix multiplication predominate gradually in a N/M^2 system as the dimension of antenna configuration scales higher. Take LLRs into consideration, the throughput can be described as follows according to different modulations, as mentioned in Literature[16],

$$Throuput=(M\cdot Q/Latency)\cdot Frequency\ , \tag{4.2}$$

in which Q indicates the modulation method. Assume 64-QAM modulation. The throughput for a 16×4 , 32×4 , 64×8 and 128×8 system is 93.8Mb/s, 130.4Mb/s, 82.2Mb/s and 107.1Mb/s, respectively. The throughput of a $N/M^2=2$ system is close to Literature[2], and is better than the $N/M^2=1$ systems.

5. Conclusion

In this paper, we firstly analyzed the computational and BER performance of 2-terms Neumann series approximation and LDL decomposition matrix inversion, and then provided a method to measure their performance gap. On this basis, we proposed a novel approximate-exact matrix inversion selection method in order to meet the balance of computational complexity and performance. This method enables the selection of an approximate inversion and an exact inversion according to different dimensions of antenna configurations relying on the proposed "critical point" of N/M^2 as estimated. Then we have implemented this selection method to our Reconfigurable Computing System as to the high dimension problems, and achieved sub-optimal results close to the literature for $N/M^2=2$ situations.

References

[1] S. Sesia, I. Toufik, and M. Baker, *LTE: The UMTS Long Term Evolution: From Theory to Practice*. Wiley Online Library pp. 1-98(2009)

[2] 3rd Generation Partnership Project; *Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA)*; Multiplexing and channel coding (Release 9). 3GPP Organizational Partners TS 36.212 Rev. 8.3.0 pp.1-21(May 2008).

[3] 3rd Generation Partnership Project; *Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (EUTRA)*; Physical Layer Procedures (Release 10). 3GPP Organizational Partners TS 36.213 version 10.10.0 pp.1-54(Jul. 2013).

[4] IEEE Draft Standard Part 11: *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications:* Amendment 4: Enhancements for Higher Throughput. P802.11n D3.00 pp.2-4(Sep. 2007).

[5] A. Paulraj, R. Nabar, and D. Gore, *Introduction to Space-Time Wireless Communications*. New York, USA: Cambridge University Press pp.56-62(2008).

[6] 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*, arXiv preprint: 1201.3210v1(Jan. 2012).

[7] T. L. Marzetta, *Noncooperative cellular wireless with unlimited numbers of base station antennas*, IEEE Trans. Wireless Commun., vol. 9, no. 11, pp. 3590–3600(Nov. 2010).

[8] E. G. Larsson , F. Tufvesson , O. Edfors and T. L. Marzetta, *Massive MIMO for next generationwireless systems*, IEEE Commun. Mag., vol. 52, no. 2, pp.186 -195(2014).

[9] J. G. Andrews, S. Buzzi, W. Choi, S. Hanly, A. C. K. Soong, J. C. Zhang, *What will 5G be?*, IEEE J. Sel. Areas Commun., vol. 32, no. 6, pp.1065 -1082(2014).

[10] R. Bouml;hnke , D. Wuuml;bben , V. Kuuml;hn and K. D. Kammeyer *Reduced complexity MMSE detection for BLAST architectures*, Proc. IEEE Global Telecommunications Conf., San Francisco, CA, USA, vol. 4, pp.2258 -2262(2003).

[11] E.G. Larsson, *MIMO detection methods: How they work [lecture notes]*. IEEE Signal Process. Mag., vol.26, no. 3 pp. 91-95(2009).

[12] B. M. Hochwald and S. ten Brink, *Achieving near-capacity on a multiple-antenna channel*, IEEE Trans. Commun., vol. 51, no. 3, pp. 389–399(2003).

[13] G.W. Stewart, Matrix Algorithms, *Vol. 1: Basic Decompositions.* Society for Industrial and Applied Mathematics (SIAM), pp.63-82(1998).

[14] M. Wu, B. Yin, A. Vosoughi, C. Studer, J. R. Cavallaro, and C. Dick, *Approximate matrix inversion for high-throughput data detection in the large-scale MIMO uplink*, in Proc. IEEE ISCAS, Beijing, China, pp. 2155–2158(May 2013)

[15] B. Yin, M. Wu, C. Studer, J. R. Cavallaro, and C. Dick, *Implementation trade-offs for linear detection in large-scale MIMO systems*, in Proc. IEEE ICASSP, Vancouver, BC, pp. 2679–2683(May 2013).

[16] . 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).