doi:10.3772/j.issn.1006-6748.2023.02.007

A TSE based design for MMSE and QRD of MIMO systems based on ASIP①

FENG Xuelin (冯雪林)②[*] [**] [***], SHI Jinglin[*], CHEN Yang[*], FU Yanlu[***], ZHANG Qineng[*], XIAO Feng[***]
([*] Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, P. R. China)
([**] School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, P. R. China)
([***] Beijing Sylincom Technology Co., Ltd., Beijing 100190, P. R. China)

Abstract

A Taylor series expansion (TSE) based design for minimum mean-square error (MMSE) and QR decomposition (QRD) of multi-input and multi-output (MIMO) systems is proposed based on application specific instruction set processor (ASIP), which uses TSE algorithm instead of resource-consuming reciprocal and reciprocal square root (RSR) operations. The aim is to give a high performance implementation for MMSE and QRD in one programmable platform simultaneously. Furthermore, instruction set architecture (ISA) and the allocation of data paths in single instruction multiple data-very long instruction word (SIMD-VLIW) architecture are provided, offering more data parallelism and instruction parallelism for different dimension matrices and operation types. Meanwhile, multiple level numerical precision can be achieved with flexible table size and expansion order in TSE ISA. The ASIP has been implemented to a 28 nm CMOS process and frequency reaches 800 MHz. Experimental results show that the proposed design provides perfect numerical precision within the fixed bit-width of the ASIP, higher matrix processing rate better than the requirements of 5G system and more rate-area efficiency comparable with ASIC implementations.

Key words: multi-input and multi-output (MIMO), minimum mean-square error (MMSE), QR decomposition (QRD), Taylor series expansion (TSE), application specific instruction set processor (ASIP), instruction set architecture (ISA), single instruction multiple data (SIMD), very long instruction word (VLIW)

0 Introduction

Coexistence of multiple wireless communication standards[1-2] and the rule of standards refreshed every 7 – 10 a have motivated the use of programmable platforms[3-8] for baseband processing (BP). Multi-input and multi-output (MIMO)[9-10] is the key technology of wideband communication systems. Thus, high-performance design for the family of representative MIMO detection algorithms[11-12] in one platform, including the classic linear minimum mean-square error (MMSE) algorithm and QR decomposition (QRD) based ML variants[13], can facilitate system to improve performance via appropriate algorithm in different scenarios, where the design and implementation of MMSE and QRD become the critical tasks.

The flexible implementations for MMSE are usually based on the application specific instruction set processor (ASIP) with architectures of single instruction multiple data (SIMD)[14] and very long instruction word (VLIW)[15-16], or coarse-grained reconfigurable architecture (CGRA)[17], which improves the matrix processing rate at large cost of area. The rates of implementations above are difficult to meet the needs of wideband BP detection, and area-efficiency is still not comparable with ASIC implementations such as Ref. [18].

Related studies on the QRD architecture can be classified into three major categories on different algorithms, such as modified Gram-Schmidt (MGS)[19-21], Givens rotation (GR)[22-23], and householder (HH)[24]. As to the flexible implementations, Ref. [19] was based on reconfigurable processing unit by bits setting and sequence flow control to get limited flexibility, while Ref. [23] used GR algorithm with longer clock latency in iteration operations and higher throughput is achieved by

① Supported by the Industrial Internet Innovation and Development Project of Ministry of Industry and Information Technology (No. GHBJ2004).
② To whom correspondence should be addressed. E-mail: fengxuelin@sylincom.com.
Received on Aug. 10, 2022

using more vector processing units (VCU).

To support QRD and MMSE simultaneously, Ref. [22] introduced an integrated circuit (IC) implementation for multitude MIMO detection algorithms by reusing QRD, aiming at energy efficiency, but resulting in poor rate and area efficiency compared with ASIC implementations in Refs [20, 21].

Meanwhile, for the numerical precision of the matrix processing, floating-point arithmetic or wider word length of fixed-point data type is used at the cost of more hardware resources consuming and additional arithmetic units design.

In order to solve the above problems, a Taylor series expansion (TSE) based design for MMSE and QRD of MIMO systems based on ASIP is proposed, with following characteristics.

(1) TSE algorithm is proposed instead of resource-consuming reciprocal and reciprocal square root (RSR) operations, to give a high performance implementation for MMSE and QRD in one programmable platform simultaneously.

(2) ISA and allocation in SIMD-VLIW architecture for MMSE and QRD are provided, which can obtain efficient data parallelism and instruction parallelism for different dimension matrices and operation types.

(3) Perfect numerical precision can be achieved with flexible table size and expansion order in TSE ISA within the fixed bit-width of the ASIP.

1 System model

Consider a MIMO system with N_T transmit and N_R receive antennas, the received symbol vector $\boldsymbol{y}$ is given by

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

where, $\boldsymbol{H} \in \boldsymbol{C}_{N_R \times N_T}$ is the complex channel matrix, $\boldsymbol{x} \in \boldsymbol{C}_{N_R \times 1}$ is the transmitted symbol vector, and $\boldsymbol{n} \in \boldsymbol{C}_{N_R \times 1}$ is the additive Gaussian noise with zero mean and variance σ_n^2.

1.1 Algorithms of matrix inversion and MMSE

The estimated transmit symbol vector $\hat{\boldsymbol{x}}$ using the MMSE algorithm is computed as shown in Eq. (2).

$$\hat{\boldsymbol{x}} = \boldsymbol{W}\boldsymbol{y},\ \boldsymbol{W} = (\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H} + \sigma_n^2\boldsymbol{I})^{-1}\boldsymbol{H}^{\mathrm{H}} \tag{2}$$

where, $\boldsymbol{W}$ is the pre-calculation of a coefficient matrix and $\boldsymbol{I}$ is the identity matrix. $(\boldsymbol{A})^{\mathrm{H}}$ is the conjugate transpose of matrix $\boldsymbol{A}$.

The method called block-wise analytic matrix inversion (BAMI) is proposed to compute the inverse of complex-valued matrices by partitioning the matrix into smaller matrices, and then compute the inverse based on computations on these smaller parts.

For example, to compute the inverse of a 4×4 matrix $\boldsymbol{M}$ in MMSE algorithm, it is divided into four submatrices $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{D}$, which can be expressed as

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} \tag{3}$$

where, $\boldsymbol{B} = \boldsymbol{C}^{\mathrm{H}}$ and the matrices of $\boldsymbol{A}$, $\boldsymbol{D}$, $\boldsymbol{M}$ are Hermitian matrices. The inversion of Eq. (3) can be described as

$$\boldsymbol{E} = \boldsymbol{M}^{-1} = \begin{bmatrix} \boldsymbol{A}^{-1} + \boldsymbol{P}^{\mathrm{H}}\boldsymbol{Q}^{-1}\boldsymbol{P} & -(\boldsymbol{Q}^{-1}\boldsymbol{P})^{\mathrm{H}} \\ -\boldsymbol{Q}^{-1}\boldsymbol{P} & \boldsymbol{Q}^{-1} \end{bmatrix} \tag{4}$$

where, $\boldsymbol{P} = \boldsymbol{C}\boldsymbol{A}^{-1}$, $\boldsymbol{P}^{\mathrm{H}} = \boldsymbol{A}^{-1}\boldsymbol{B}$, $\boldsymbol{Q} = \boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}$.

The inversion of a 2×2 matrix can be computed as

$$\boldsymbol{A}^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}^{-1} \tag{5}$$

where the term of $(ad - bc)$ in Hermitian matrix is a real value.

The architecture of MMSE matrix inversion is depicted in Fig. 1, which shows five kinds matrix operations. One Hermit matrix inversion can be obtained by two submatrices inverse as 'Inv', four submatrices multiplication as 'Mul', single submatrix addition as 'Add', single submatrix subtraction as 'Sub', and two matrices conjugate transpose as '$(\ )^{\mathrm{H}}$'.

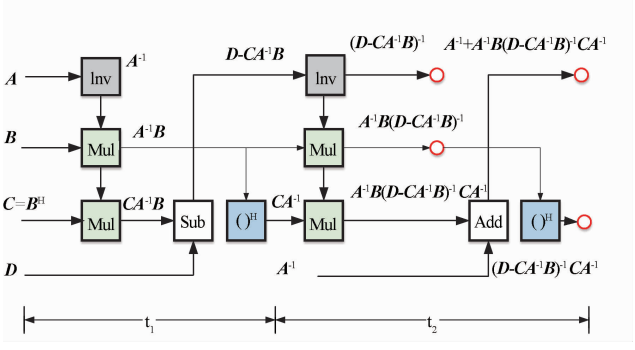

Fig. 1 Matrix inversion architecture

1.2 Algorithms of MGS-QRD

The MGS-QRD decomposes an $n \times n$ matrix $\boldsymbol{H}$ into an $n \times n$ unitary matrix $\boldsymbol{Q}$ and an $n \times n$ upper triangular matrix $\boldsymbol{R}$. For simplicity, a 4×4 complex matrix is utilized to explain the procedure as Eq. (6).

$$\begin{cases} \boldsymbol{DP}: \begin{cases} r_{ii} = \|\boldsymbol{a}_i^k\|_2 \\ \boldsymbol{q}_i = \boldsymbol{a}_i^k / r_{ii} \end{cases}, i \in \{1, 2, 3, 4\},\ k = i \\ \boldsymbol{TP}: \begin{cases} r_{ij} = (\boldsymbol{q}_i)^{\boldsymbol{H}}\boldsymbol{a}_j^k \\ \boldsymbol{a}_j^{k+1} = \boldsymbol{a}_j^k - r_{ij}\boldsymbol{q}_i \end{cases}, j \in \{1, 2, 3, 4\},\ j > i,\ k = i \end{cases} \tag{6}$$

where the classical MGS algorithm is composed of 4 iterations, and the QRD can be separated into two steps. $\boldsymbol{a}_i^k$ denotes the column i of matrix $\boldsymbol{H}$ in the kth iteration. $\boldsymbol{q}_j$ represents the column j of matrix $\boldsymbol{Q}$ and $\boldsymbol{r}_{ij}$ denotes row i and column j of matrix $\boldsymbol{R}$. $\|v\|_2$ means the norm of a vector $\boldsymbol{v}$.

The first step diagonal-process (DP), computes the diagonal line elements of the upper triangular matrix $\boldsymbol{R}$ and the column of the unitary matrix $\boldsymbol{Q}$. The second step triangular-process (TP), computes the non-diagonal line elements of matrix $\boldsymbol{R}$ and renew the column elements of matrix $\boldsymbol{H}$.

The architecture of QRD is depicted in Fig. 2, which shows that one matrix QRD can be carried out by 7 stages, including 4 DP stages and 3 TP stages. The TP stages finish 6 TP units in total to calculate the non-diagonal line elements of matrix $\boldsymbol{R}$, where $\boldsymbol{TP}_{jk}$ denotes the jth $\boldsymbol{TP}$ operation of the kth iteration.

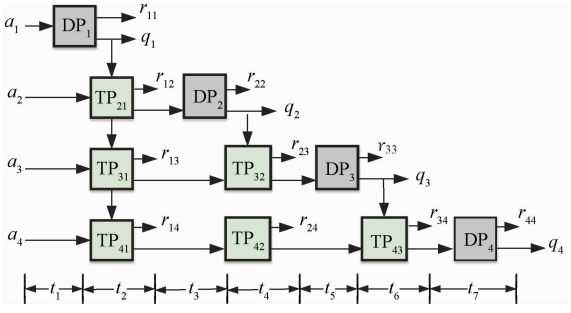

Fig. 2 MGS-QRD architecture

1.3 Implementation demands and BP-ASIP architecture

As to 5G communication system of 100 MHz bandwidth, 3300 subcarriers are used to transmit the 4 × 4 MIMO OFDM symbols in each TTI of 0.5 μs, which means the ratio of 4 × 4 matrix inversion or QRD should be up to about 79.2M matrix/s.

This work implements the algorithms of MMSE and QRD on a BP-ASIP proposed in Ref. [7], whose architecture is VLIW and SIMD mixed. The BP-ASIP has 512-bit SIMD data paths and 192-bit VLIW instruction word length, supporting 32-lane 16-bit or 16-lane 32-bit computing and 6 instructions dispatch respectively.

Table 1 shows the VLIW architecture based on the kernel algorithms of baseband processing: vector computing (VCMU), vector shuffle/alignment (VSHF), vector load/store (VL/S), program flow control (SALU), vector/scalar data exchanging (SVEX), and address generation (AGU).

Table 1 VLIW architecture of BP-ASIP

VLIW#0	VLIW #1	VLIW#2	VLIW#3	VLIW#4	VLIW#5
VCMU	VSHF	VL/S	SALU	SVEX	AGU

The ASIP has nine-stage pipelines to achieve high frequency as Fig. 3, four-pipeline fetch operation (FE), single-pipeline instruction decode (DC), three-pipeline execution (EX) including MEM load/store, and single-pipeline written back (WB) to register. Based on the pipeline technique, five instructions can be implemented for uncorrelated data in every five clocks from DC to WB.

Fig. 3 Nine-stage pipelines of BP-ASIP

2 Algorithm and insructions

Reciprocal operation and RSR is the key of matrix inversion of MMSE algorithm and QRD calculation respectively. TSE algorithm and instructions are proposed to reduce the reciprocal hardware resources, and to support the RSR operations simultaneously.

Two kinds of instructions are also designed to accelerate the MMSE detection and QRD respectively.

2.1 TSE algorithm

The core idea of TSE algorithm is to compute two resource-consuming operations of reciprocal operation $1/x$ and the RSR operation $1/\sqrt{x}$ simultaneously without division operations.

The two operations can be approximated via TSE algorithm at point x_0 as Eq. (7).

$$f(x)=\begin{cases}\dfrac{1}{x}\approx\underbrace{\dfrac{1}{x_0^3}}_{a_2}x^2-\underbrace{\dfrac{3}{x_0^2}}_{a_1}x+\underbrace{\dfrac{3}{x_0}}_{a_0} & x_0\in(0.5,2]\\ \dfrac{1}{\sqrt{x}}\approx\underbrace{\dfrac{3}{8x_0^{2.5}}}_{a_2}x^2-\underbrace{\dfrac{5}{4x_0^{1.5}}}_{a_1}x+\underbrace{\dfrac{15}{8x_0^{0.5}}}_{a_0} & x_0\in(0.5,2]\end{cases} \tag{7}$$

where, a_2, a_1, and a_0 depend on the operation and the expansion point x_0. Thus the value set of $\{a_2, a_1, a_0\}$ corresponding to x_0 in interval can be stored in a look-up table (LUT). Then the value of $\{a_2, a_1, a_0\}$ can be obtained according to the point of x_0 nearest to x.

2.2 Instructions

2.2.1 Instructions for TSE, MMSE and QRD

The guidelines for instruction set design are completeness, orthogonality, regularity and simplicity, ease of programming and implementation.

Firstly, LUT instructions, matrix inversion instructions, and vector inner product instructions, are proposed for TSE, MMSE, and QRD respectively

based on their completeness characteristics. In addition, a fine-grained ISA is designed with orthogonality and is more conducive to parallel execution in VLIW architecture. The ISA including regular addition, subtraction, multiplication and conjugate transpose for vectors or matrices, are simplified and reusable as much as possible. Thus, MMSE and QRD are easy to program and implement, with different modes of TSE and the simplify instructions designed.

Table 2 shows TSE ISA proposed for the reciprocal and RSR.

Table 2 TSE ISA

Instruction	Function
(1) *v. LUTtaylor*	Get $Vt = \{a_3, a_2, a_1, a_0\}$ with x of $Vs = \{x^3, x^2, x, 1\}$ according to mask bits and mode configuration
(2) *v. mul*	Get dot product $Vd = Vs\ Vt$ according to mask bits
(3) *v. iadd*	Get $1/x$ or $1/\sqrt{x}$ by sum (*vd*) of vector inner addition

Where *v. LUTtaylor* loads the expansion series of x_0 according to the input vector by LUT operation with different mode configuration, and the value of TSE is calculated by *v. mul* and *v. iadd*, for dot product of two vectors and vector inner addition respectively.

Table 3 shows six instructions proposed for the four parts having the intermediate values of $\boldsymbol{H}^{\mathrm{H}}$, $\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H} + \sigma_n^2\mathbf{I}$, $(\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H} + \sigma_n^2\mathbf{I})^{-1}$, and $(\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H} + \sigma_n^2\mathbf{I})^{-1}\boldsymbol{H}^{\mathrm{H}}$ of MMSE detection.

Table 3 Six new instructions for MMSE

Operation	Instruction	Function
$(\,)^{\mathrm{H}}$	(4) *v. mhermit*	Get matrix conjugate transpose
MMult.	(5) *v. mmul*	Get matrix multiplication
Add.	(6) *v. add*	Get addition for vector or matrix
Sub.	(7) *v. sub*	Get subtraction for vector or matrix
2×2 Matrix Invers.	(8) *v. minv2det*	Get $Vd = \{x^3, x^2, x, 1\}$ by $Vs = \{a, b, c, d\}$, where $x = ad\text{-}bc$
	(9) *v. minv2shuf*	Get $Vt = \{d, -b, -c, a\}$ by *Vs*
	TSE, *v. mul*	Get Matrix *Inv.* by $Vt/(ad\text{-}bc)$

Where the former four instructions support for calculations for 4×4 matrix of MMSE intermediate values and 2×2 matrix of 4×4 BAMI algorithm with different mask bits. As to 2×2 matrix inversion, it can be finished by six instructions including *v. minv2det*, *v. minv2shuf*, TSE ISA of reciprocal mode and *v. mul*.

Table 4 shows two instructions proposed for DP and TP of QRD.

Table 4 Two new instructions for QRD operations

Operation	Instruction	Function
DP	(10) *v. modu2* *TSE*, *v. mul*	Get $Vs = \{x^3, x^2, x, 1\}$, where $x = \|a_i\|^2$, Get $1/r_{ii} = 1/\sqrt{x}$, q_i, and r_{ii}
TP	(11) *v. ip* *v. mul*, *v. sub*	Get r_{ij} by inner product of q_i and a_j^k Get $r_{ij} \cdot q_i$, and $a_j^{k+1} = a_j^k - r_{ij}q_i$

Where DP is implemented in five or six instructions with *v. modu2*, TSE ISA of RSR mode and *v. mul*, in which r_{ii} is not the critical path and could be calculated latter or not calculated. TP is finished in three instructions with *v. ip*, *v. mul* and *v. sub*.

The instructions proposed above are realized on the SIMD-VLIW processor BP-ASIP, and allocated as Table 5.

Table 5 ISA allocation in DATA PATHs of the ASIP

VLIW	Function Units	Instruction index
Data Path 0	VCMU	(2), (5), (6), (7), (8), (10), (11)
Data Path 1	VSHF	(3), (4), (9)
Data Path 5	AGU/LUT	(1)

2.2.2 Instruction flow of matrix inversion and QRD

For matrix multiplication and inversion, elements of matrix are all placed row by row in vector register (VR). A 16-lane 32-bit VR can process four 2×2 matrices or one 4×4 matrix.

Fig. 4 shows the instruction flow of matrix inversion, four 4×4 compleces matrices inversion can be processed in parallel according to the BAMI algorithm, and be finished within 18 instructions, thanks to the VLIW multiple data paths.

Based on the pipeline technology, one 4×4 complex matrix inversion can be carried out in 4.5 cycles on average, and one 4×4 complex $\boldsymbol{W}$ for MMSE will take 8.5 cycles in total with 4 cycles consumption for $\boldsymbol{H}^{\mathrm{H}}$, $\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H} + \sigma_n^2\mathbf{I}$ and $(\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H} + \sigma_n^2\mathbf{I})^{-1}\boldsymbol{H}^{\mathrm{H}}$.

For QRD, the calculation is implemented by column, and columns of a matrix are placed in different VRs. The 16-lane 32-bit VR can process four 4×4 matrices simultaneously.

Fig. 5 shows the instruction flow of QRD, where calculations and load/store operations, including DP4 of the last matrix and DP1 of the new matrix, are processed in parallel because of the VLIW architecture. While the number of matrices to be processed is large usually, one 4×4 complex QRD can be calculated in about 8.5 cycles

on average by the pipeline technology.

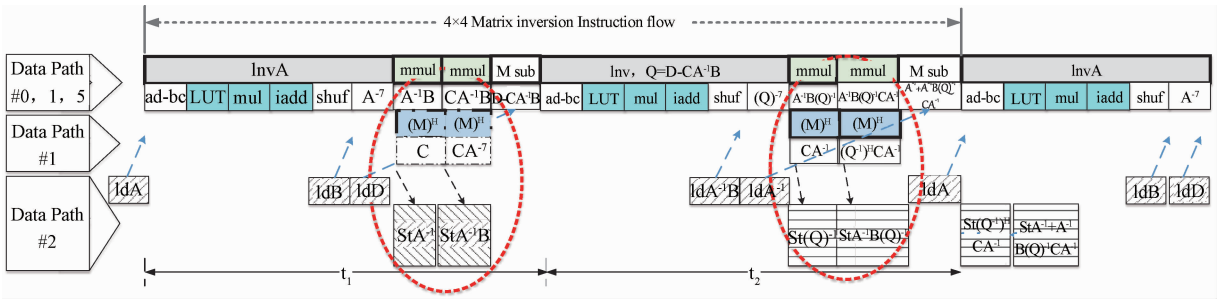

Fig. 4 ASIP 4X4 matrix inversion instruction flow

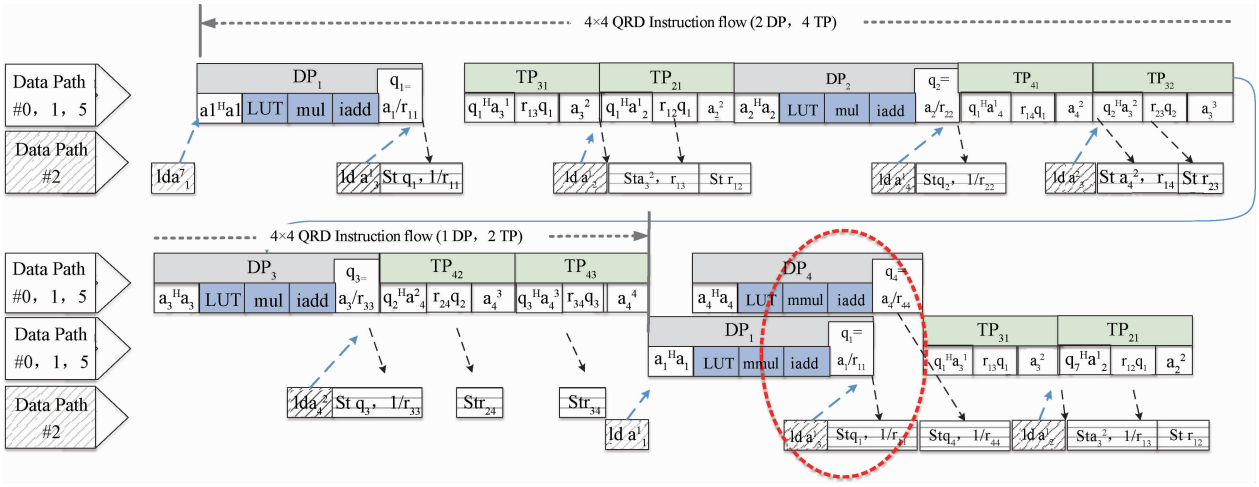

Fig. 5 ASIP 4X4 matrix QRD instructions flow

2.3 Expand to 8 ×8 MMSE and QRD

For 5G communication system, the data stream parallelism can be up to 8, which means the MMSE and QRD algorithm need to consider the 8 ×8 case.

Matrix multiplication and inversion of 8 ×8 matrix can be divided into four 4 ×4 submatrices according to Eqs(3) and (4), and can be carried out by the operations of 4 ×4 matrix.

As to the 8 × 8 QRD GMS decomposition algorithm, the number of elements participating in the inner product of vectors changes from 4 to 8, by adjusting the mask-bits in the ISA.

Therefore, the proposed algorithms and ISA not only support 2 ×2, 4 ×4 matrix operations, but also support 8 ×8 or even 16 ×16 matrix operations through the extension of the algorithm and ISA.

3 Implementation results and comparison

In order to evaluate the performance, area and power consumption of the proposed algorithm and instruction set in this paper. Performance of the TSE algorithm is evaluated with different expansion orders and fixed-point lengths, and the TSE instruction set is implemented in the BP-ASIP processor. The comparison results between this platform and other related literatures are given.

3.1 MSE performance

Fig. 6 shows the MSE of three methods for reciprocal, including expansion of second-order and third-order.

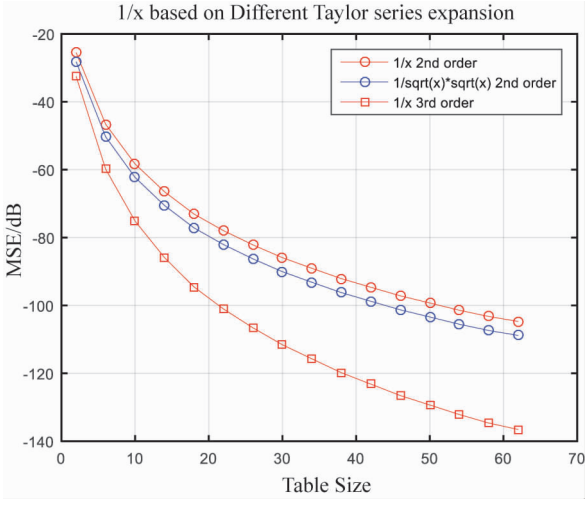

Fig. 6 MSE of reciprocal with different TSE orders

When the table size is 10, MSE of the above three methods can reach about −60 dB, which can meet the requirements of practical engineering applications. The table size and TSE order can be flexibly selected according to the precision demand.

The fixed-point MSE of TSE for different modes are shown in Fig. 7 and Fig. 8 respectively.

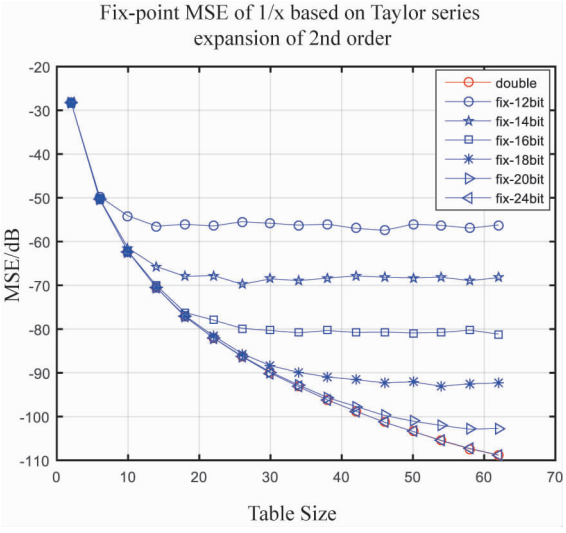

Fig. 7 MSE of TSE IS in reciprocal mode

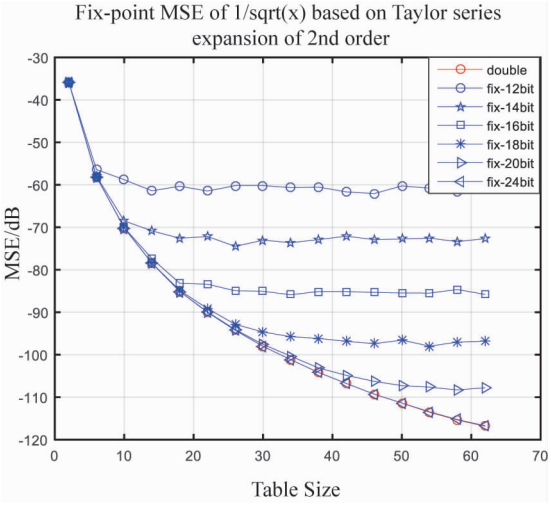

Fig. 8 MSE of TSE in RSR mode

Fig. 7 indicates that MSE of reciprocal can achieve about −60 dB with table-size of 10 and fixed-point bit-width of 14 bit, and achieve about −78 dB with table-size of 18 and fixed-point bit-width of 16 bit respectively.

Fig. 8 indicates that MSE of RSR can achieve about −68 dB with table-size of 10 and fixed-point bit-width of 14 bit, and achieve about −82 dB with table-size of 18 and fixed-point bit-width of 16 bit respectively.

When fixed-point bit-width is 24 bit, precisions of reciprocal and RS are both close to double floating-point's precisions.

3.2 Implementation performance

To evaluate the performance of the designed ISA proposed for TSE, MMSE and MGS algorithms, it is developed on the baseband processor of BP-ASIP, which also supports the other baseband processing functions of 5G. The ASIP is implemented with TSMC 28 nm CMOS technology and the work frequency is 800 MHz. Thus the processing rate of MMSE and QRD reaches 94.12M Matrix/s, which exceeds 79.2M Matrix/s, the requirements of the 5G system@100 MHz.

To fairly compare the rate of MMSE and QRD with the other existing work, the normalized rate (NR) and the normalized rate of area (NRA) are defined to evaluate the throughput performance as Eq. (8) and Eq. (9).

$$NR = Rate \cdot \frac{tech}{28\ \text{nm}} \cdot \frac{1}{\eta} \cdot \frac{16}{SIMD} \cdot \frac{1}{VCU} \tag{8}$$

$$NRA = Rate \cdot \frac{tech}{28\ \text{nm}} \cdot \frac{1}{\eta} \cdot \frac{1}{Area} \tag{9}$$

where the *Rate* denotes the number of matrices processed per second. Additionally, the technology parameter *tech* is normalized to 28 nm CMOS technology. When the matrix is a real matrix, the workload parameter η is 4, otherwise η is 1. *SIMD* denotes the number of data units be processed in parallel, and *VCU* is the number of the VCU. *Area* is the area of each implementation in k gate counts.

Table 6 illustrates the comparison results of this work with numerous existing MMSE and QRD work including ASIP and ASIC implementations.

As to ASIP implementations, NR and NRA of MMSE and QRD in this work outperform most of the other literatures.

In the case of MMSE, NR and NRA performances of this work are 12.3% and 69.6% better than the average ones of Refs[14-16] respectively. Compared with Ref. [15], NR performance is less, but NRA is 2.73 times better than that. Analyzing the differences of the implementation literatures, Refs[14-16] are all based on BAMI algorithm, which used SIMD or SIMD-VLIW architecture to increase the data parallelism, but did not take advantage of bit-width when the data length of different operations varied. In addition, Ref. [16] adopted floating-point design, consuming large cycle count and power. Ref. [17] adopted iterative algorithm of division-free ShermonMorries to use regular CGRA PE with function units and local register files, consuming large area resources. In the case of QRD, NR and NRA of this work is close to 35% better than Ref. [19]

respectively, which used regular PEs to process the entire MGS flow and did not use pipeline to speed up the process. NRA of this work is 45% better than Ref. [23], which produced performance with more VCU units to compensate the large delay of GR algorithm.

Compared with the ASIP literatures above, ISA of this work is more flexible and gets used effectively for different dimension matrices and operation types. Meanwhile, based on the ISA and bit-width, MSE can reach the same magnitude and be lower with third-order TSE. Furthermore, the ISA is more fine-grained, which makes the pipeline technology and VLIW architecture utilized more effectively.

As to ASIC implementations, compared with the hardware resources-saving schemes, NR and NRA of this work are comparable with Refs [18] and [20], which used iterative design to reduce the resources and led to lower throughput. Compared with the high throughput scheme, NRA of this work is only 45% less than Ref. [21], which had high pipelined units and dedicated design for each operation. Finally, NRA of this work is far better than the IC implementation[22], which supports multitude MIMO detection algorithms by reusing QRD, aiming at energy efficiency, but resulting in poor rate and area efficiency.

It is clear that this work has an advantage to support higher performance MMSE and QRD simultaneously, and the advantage will become even greater considering more requirements for programmable baseband processing units of the future communication systems.

Table 6 Comparison results with Refs[14-23]

	MMSE Flexible Imp Lementations					MMSE ASIC	QRD Flexible Imp Lementations			QRDASIC		Multi-mode	
References	This	[14]	[15]	[16]	[17]	[18]	This	[19]	[23]	[20]	[21]	[22]	
Algorithm	BAMI	BAMI	BAMI	BAMI	S. M	cholesky	CMGS	CMGS	CGR	RMGS	RMGS	BAMI	CGR
Format	Fi-16	FL-16	FL-16	FL-18	Fi-22	BFP-23	Fi-16	Fi-24	Fi&FL	Fi-23	Fi-16	Fi-16	
Cycles	8.5	202/4	166/2	69	17	16	8.5	24	5	35	3	58	20
Frequency/MHz	800	400	250	400	263	453	800	256	1300	400	416	166	
Tech/nm	28	65	90	90	65	90	28	65	28	180	130	65	
Rate (M M/s)	94.12	7.93	3	5.8	15.47	28.3	94.12	10.66	260	11.5	138.67	2.86	8.3
Area(K Gc)	420	90	210 *	123	597	299	420	149	-	32.6	328	469 *	
Power(mW)	83	N. M	-	200	-	-	83	48.2	-	15.5	-	300.9	
SIMD, VCU(unit)	16, 1	4, 1	2, 1	4, 1	1, 1	-	16, 1	-	16, 4	-	-	-	-
NR (M M/s/unit)	94.12	73.6	77.1	74.6	574.6	90.96	94.12	-	65	18.5	160.9	6.6	19.3
NRA (MM/s/kGc)	0.224	0.205	0.046	0.152	0.06	0.304	0.224	0.166	-	0.567	0.491	0.014	0.041

* Excluding the area of memory

4 Conclusion

A Taylor series expansion algorithm based design, for MMSE and QRD of MIMO systems based on ASIP, is proposed. TSE algorithm is used to replace two resource-consuming operations of reciprocal and RSR, and to support MMSE and QRD simultaneously for MIMO baseband receiver. Furthermore, instruction set and the allocation in SIMD-VLIW architecture for MMSE and QRD are provided, which can obtain efficient data parallelism and instruction parallelism for different dimension matrices and operation types. Meanwhile, perfect numerical precision can be achieved using the TSE ISA of table lookup and vector inner product function, which supports flexible table size and TSE order. Experimental results show that the proposed algorithm provides higher matrix throughput better than the requirements of 5G system, more rate-area efficiency comparable with the ASIC implementations, and multiple magnitude of precision within the bit-width of ASIP data type. In future research, higher rate-area efficiency implementation will be considered aiming at both flexibility and efficiency for ASIP.

References

[1] Cisco. Cisco annual internet report (2018 – 2023) white paper[EB/OL]. [2022-11-01]. https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html.

[2] LIU L, ZHOU Y, YUAN J, et al. Economically optimal MS association for multimedia content delivery in cache-enabled heterogeneous cloud radio access networks[J]. IEEE Journal on Selected Areas in Communications, 2019, 37:1584-1593.

[3] ANJUM O, AHONEN T, GARZIA F, et al. State of the art baseband DSP platforms for software defined radio: a survey[J]. EURASIP Journal on Wireless Communications and Networking, 2011(5):5-24.

[4] LI L. IC challenges in 5G[C]//2015 IEEE Asian Solid-

State Circuits Conference (A-SSCC). Xiamen: IEEE, 2016:1-4.
[5] HANDAGALA S, LEESER M. Real time receiver baseband processing platform for sub 6 GHz PHY layer experiments[J]. IEEE Access, 2020(8):105571-105586.
[6] CHEN Y, LIU L, FENG X, et al. DXT501: an SDR-based baseband MP-SoC for multi-protocol industrial wireless communication[C]//2022 IEEE Symposium in Low-Power and High-Speed Chips (COOL CHIPS). Tokyo: IEEE, 2022:1-6.
[7] ZHU Z, SHAN T, SU Y, et al. A 100 GOPS ASP based baseband processor for wireless communication[C]//Design, Automation & Test in Europe Conference & Exhibition. Grenoble: IEEE, 2013:121-124.
[8] YAO Y, SU Y, SHI J, et al. A low-complexity soft QAM de-mapper based on first-order linear approximation[C]//IEEE International Symposium on Personal. Hong Kong: IEEE, 2015:446-450.
[9] PAULRAJ A J, GORE D A, NABAR R U, et al. An overview of MIMO communications—a key to gigabit wireless[J]. Proceedings of the IEEE, 2004, 92(2): 198-218.
[10] SHAFI M, MOLISCH A F, SMITH P J, et al. 5G: a tutorial overview of standards, trials, challenges, deployment and practice[J]. IEEE Journal on Selected Areas in Communications, 2017, 35(6):1201-1221.
[11] LARSSON E G. MIMO detection methods: how they work [J]. IEEE Signal Processing Magazine, 2009, 26(3): 91-95.
[12] YANG S, HANZO L. Fifty years of MIMO detection: the road to large-scale MIMOs[J]. IEEE Communications Surveys & Tutorials, 2015, 17(4):1941-1988.
[13] MIN L, BOUGARD B, LOPEZ E E, et al. Selective spanning with fast enumeration: a near maximum-likelihood MIMO detector designed for parallel programmable baseband architectures[C]//IEEE International Conference on Communications. Beijing: IEEE, 2008:737-741.
[14] EILERT J, DI W, LIU D. Implementation of a programmable linear MMSE detector for MIMO-OFDM[C]//International Conference on Acoustics, Speech, and Signal Processing. Las Vegas: IEEE, 2008:5396-5399.
[15] EBERLI S, CESCATO D, FICHTNER W. Divide-and-conquer matrix inversion for linear MMSE detection in SDR MIMO receivers[C]//NORCHIP 2008. Tallinn: IEEE, 2008:162-167.
[16] GUENTHER D, LEUPERS R, ASCHEID R. Efficiency enablers of lightweight SDR for MIMO baseband processing[J]. IEEE Transactions on Very Large scale Integration (VLSI) Systems, 2016, 24(2):567-577.
[17] CHEN X, MINWEGEN A, HASSAN Y, et al. FLEXDET: flexible, efficient multi-mode MIMO detection using reconfigurable ASIP[C]//IEEE 20th International Symposium on Field-Programmable Custom Computing Machines. Toronto: IEEE, 2012:69-76.
[18] SENNING C, BURG A. Block-floating-point enhanced MMSE filter matrix computation for MIMO-OFDM communication systems[C]//IEEE International Conference on Electronics. Abu Dhabi: IEEE, 2013:787-790.
[19] PRADHAN A K, Nandy S K. An energy efficient dynamically reconfigurable QR decomposition for wireless MIMO communication[C]//International Conference on VLSI Design & International Conference on Embedded Systems. Kolkata: IEEE, 2016:276-281.
[20] CHANG R C, LIN C, LIN K, et al. Iterative QR decomposition architecture using the modified Gram-Schmidt algorithm for MIMO systems[J]. IEEE Transactions on Circuits & Systems Part I Regular Papers, 2010, 57(5): 1095-1102.
[21] CHEN L, LIU C, YU W, et al. Low latency QRD algorithm for future communication[J]. IEICE Electronics Express, 2017, 14(22):1-12.
[22] MOHAMED M I A, MOHAMMED K, DANESHRAD B. Energy efficient programmable MIMO decoder accelerator chip in 65-nm CMOS[J]. Very Large Scale Integration (VLSI) Systems, 2014, 22(7):1481-1490.
[23] CEVA. CEVA PentaG UE architecture introduction[R]. 2019.
[24] KURNIAWAN I H, YOON J H, PARK J. Multidimensional householder based high-speed QR decomposition architecture for MIMO receivers[C]//IEEE International Symposium on Circuits & Systems. Beijing: IEEE, 2013: 2159-2162.

FENG Xuelin, born in 1983. She is a senior engineer of Institute of Computing Technology and Ph. D candidate at Computer Science and Technology, Chinese Academy of Sciences. She received her M. S. degree in 2009 and B. S. degree in 2006 from Beijing University of Posts and Telecommunications. Her research interests include broadband wireless signal processing, wireless terminal baseband chip algorithm design, and processor design.