A VLSI Design of an Arrayed Pipelined Tomlinson-Harashima Precoder for MU-MIMO Systems

Kosuke Shimazaki*, Shingo Yoshizawa†, Yasuyuki Hatakawa‡, Tomoko Matsumoto‡, Satoshi Konishi‡, and Yoshikazu Miyanaga*

*Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan

†Department of Electrical and Electronic Engineering, Kitami Institute of Technology, Kitami, Japan

‡KDDI R&D Laboratories Inc. Fujimino, Japan

***Abstract*—This paper presents a VLSI design of a Tomlinson-Harashima (TH) precoder for multi-user MIMO (MU-MIMO) systems. The TH precoder consists of LQ decomposition (LQD), interference cancellation (IC), and weight coefficient multiplication (WCM) units. The LQ decomposition unit is based on an application specific instruction-set processor (ASIP) architecture with floating-point arithmetic for high accuracy operations. As for the IC and WCM units with fixed-point arithmetic, the proposed architecture keeps calculation accuracy and gives shorter pipeline latency and smaller circuit size by employing an arrayed structure. The implementation result shows that the proposed architecture reduces circuit area and power consumption by 11% and 15%, respectively.**

I. Introduction

Multiple-input multiple-output (MIMO) systems are attracting attention. MIMO is a technique to improve speed and capacity in wireless communication by increasing transmit and receive antennas, which is adopted in wireless LAN standard of IEEE 802.11n. The next-generation standard of IEEE 802.11ac is being formulated for achievement of more than 1 Gbps throughputs and supports multi-user MIMO (MU-MIMO) for communication capacity enhancement. MU-MIMO has larger communication capacity than single-user MIMO (SU-MIMO) by using parallel transmission among multiple terminals. MU-MIMO requires precoding at the transmitter side for interference cancellation among receiving terminals. Types of MU-MIMO systems are classified into linear and non-linear precoding schemes. Linear precoding by zero-forcing (ZF) and minimum mean square error (MMSE) methods just multiplies transmitting signals by precoding weights [1], [2]. The communication quality tends to be degraded under high spatial correlations among users. Non-linear precoding by vector perturbation (VP) and Tomlinson-Harashima precoding (THP) reduces transmission power by coding signals and overcomes the weakness in linear precoding [3], [4]. The non-linear precoding has better communication quality, however requires larger computational complexity.

Hardware architectures of THP have been presented in [5], [6]. Their architectures focus on single-input and single-output (SISO) systems. THP testbed used a digital signal processor (DSP) for MU-MIMO has developed in [7], but the testbed computes using a lot of DSP units and power consumption and efficiency of LSI circuit area have not been discussed. We present a TH precoder consisting of LQ decomposition, interference cancellation (IC), and weight coefficient multiplication (WCM) units. The LQ decomposition unit is based on an ASIP architecture with floating-point arithmetic for high accuracy operations. We refer the ASIP for singular value decomposition (SVD) designed in our previous work [8]. As for the IC and WCM units with fixed-point arithmetic, the conventional architecture using straightforward computation takes many pipeline latency cycles in successive interference cancellation of MIMO-THP. The proposed architecture makes use of pre-computation in the LQ decomposition unit and adopts an arrayed structure, which achieves shorter pipeline latency than the conventional architecture. In the VLSI implementation, we indicate that the proposed architecture decreases processing latency time and power consumption.

This paper is organized as follows. We explain theory of THP in Section II. Section III presents a design of the THP circuit. Section IV describes performance evaluation of the designed circuit. Section V concludes the paper.

II. Tomlinson-Harashima precoding for MU-MIMO Systems

We assume a MU-MIMO system with four transmit antennas and two double-antenna users, i.e., 4x2MU-MIMO and that channel state information (CSI) is ideally fed back from a receiver to a transmitter. In THP, signals are transmitted after subtracting multi-user interference caused by the propagation channel. The channel matrix $\boldsymbol{H}$ is estimated in a receiver and decomposed into a lower triangular matrix $\boldsymbol{L}$ and a unitary matrix $\boldsymbol{Q}$.

$$\begin{aligned}\boldsymbol{H} &= \boldsymbol{LQ} \\ &= \begin{bmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} \end{bmatrix} \begin{bmatrix} q_{11} & q_{12} & q_{13} & q_{14} \\ q_{21} & q_{22} & q_{23} & q_{24} \\ q_{31} & q_{32} & q_{33} & q_{34} \\ q_{41} & q_{42} & q_{43} & q_{44} \end{bmatrix}. \end{aligned} \quad (1)$$

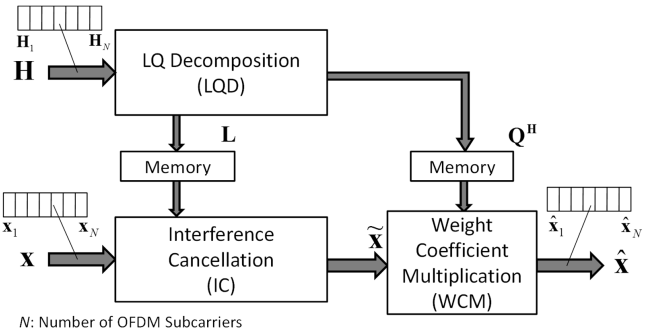


Fig. 1. Overall structure of TH precoder.

The transmitted signals $\tilde{\boldsymbol{x}}$ are generated using l_{ij} in (1) as

$$\begin{cases} \tilde{x}_1 = Mod\left(x_1\right) = x_1 \\ \tilde{x}_2 = Mod\left(x_2 - \frac{l_{21}}{l_{22}}\tilde{x}_1\right) \\ \tilde{x}_3 = Mod\left(x_3 - \frac{l_{31}}{l_{33}}\tilde{x}_1 - \frac{l_{32}}{l_{33}}\tilde{x}_2\right) \\ \tilde{x}_4 = Mod\left(x_4 - \frac{l_{41}}{l_{44}}\tilde{x}_1 - \frac{l_{42}}{l_{44}}\tilde{x}_2 - \frac{l_{43}}{l_{44}}\tilde{x}_3\right). \end{cases} \tag{2}$$

The signals are transmitted from each antenna after weight coefficient multiplication by $\boldsymbol{W}$=$\boldsymbol{Q}^H$. The received signals $\tilde{\boldsymbol{y}}$ through the propagation channel are expressed as

$$\begin{aligned} \tilde{\boldsymbol{y}} &= \boldsymbol{HW}\tilde{\boldsymbol{x}} + \boldsymbol{n} \\ &= \boldsymbol{LQQ}^H\tilde{\boldsymbol{x}} + \boldsymbol{n} \\ &= \boldsymbol{L}\tilde{\boldsymbol{x}} + \boldsymbol{n}, \end{aligned} \tag{3}$$

where $\boldsymbol{n}$ is a white Gaussian noise. Here, modulo operation in (2) is denoted as

$$Mod\left(x\right) = x - floor\left(\frac{x + M + jM}{2M}\right) \times 2M, \tag{4}$$

where M is a modulo window size. The purpose of modulo operation is to suppress amplitude of signals by transforming signals into $[-M,M]$. The amplitude of signals increases depending on l_{ij}. Power efficiency of the transmitted signals becomes worse if the signals increase by multiplying l_{ij}. A modulo window size depends on modulation levels [2].

III. VLSI Design

A. Overall structure

The TH precoder consists of the LQ decomposition (LQD), interference cancellation (IC) and weight coefficient multiplication (WCM) units, which are illustrated in Fig.1. Their requirements for calculation accuracy and throughput performance are not identical. The throughput requirement of LQD unit is not high because the CSI does not change frequently. However, the LQD unit requests high accuracy operations in executing the matrix decomposition. On the other hand, the throughput requirement of the IC and the WCM units are high.

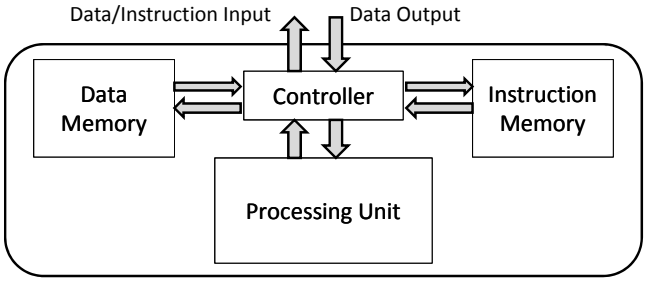


Fig. 2. Circuit structure of ASIP.

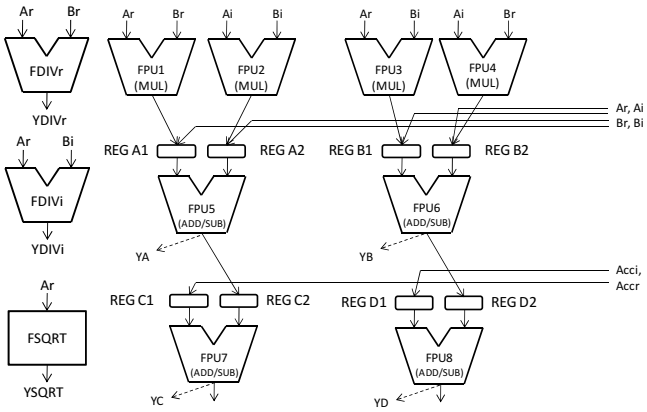


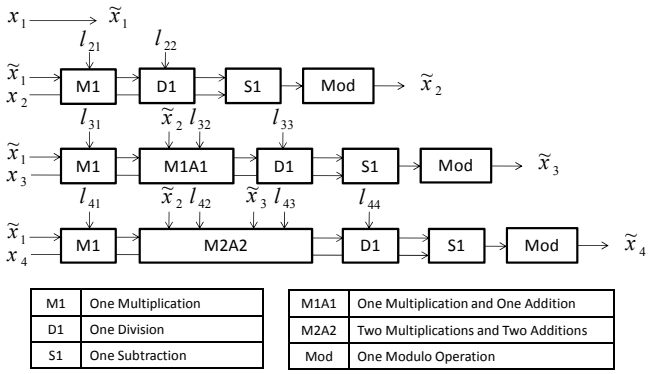
Fig. 3. Circuit structure of processing unit.

B. LQD unit

The LQ decomposition unit is based on an ASIP architecture, which is illustrated in Fig. 2. Data and instructions are stored in each memory and the processing unit executes instructions in order. We use floating-point units (FPUs) dealing with IEEE 754 standard single precision floating-point in the processing unit. The circuit structure of the processing unit is illustrated in Fig. 3. The FPU supports four types of arithmetic operations of addition, subtraction, multiplication, and division. The processing unit can execute complicated processing such as complex and accumulative operations by combining the eight FPUs. The four FPUs in the first stage are used for one complex and two real multiplications. By combining these FPUs, complex multiplication can be executed. We also design the dedicated high-speed division and square-root operation units denoted by "FDIV" and "FQRT". The instruction sets and hardware processing for matrix decomposition (Gram-Schmidt algorithm) has been presented in [8].

C. IC and WCM units

1) Conventional architecture: The conventional architecture which we assume performs straightforward computation of (2). The structure of the IC unit is shown in Fig. 4. The IC unit has operation blocks such as "M1" and "D1". These blocks have one or two pipeline stages for complex arithmetic operations with real and imaginary data. The detailed structure of the modulo operation block is denoted in Fig. 5. The mod-



M1	One Multiplication
D1	One Division
S1	One Subtraction

M1A1	One Multiplication and One Addition
M2A2	Two Multiplications and Two Additions
Mod	One Modulo Operation

Fig. 4. Conventional architecture in IC unit.

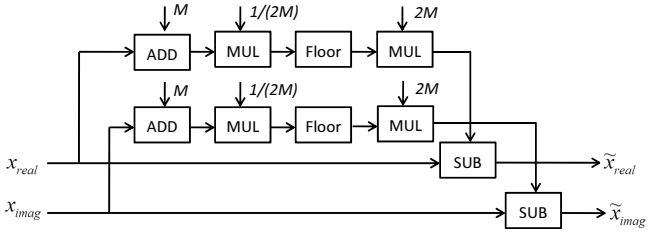
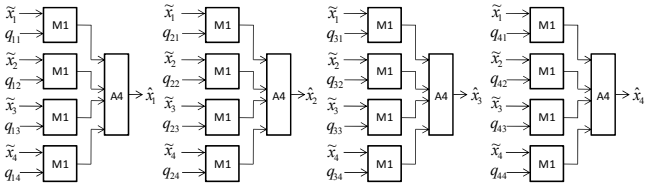
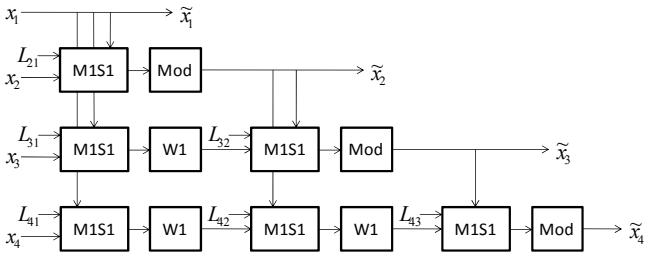


Fig. 5. Modulo operation block.



M1	One Multiplication
A4	Four Additions

Fig. 6. Conventional architecture in WCM unit.



M1S1	One Multiplication and One Subtraction
W1	One Wait
Mod	One Modulo Operation

Fig. 7. Proposed architecture in IC unit.

ulo block performs arithmetic and floor operations according to (4). The conventional architecture in the WCM units is illustrated in Fig. 6. The matrix operation of $\boldsymbol{Q}^H\tilde{\boldsymbol{x}}$ is done by complex multiplication and additions. The block of "A4" has four input ports and generates result data by one output port.

We point out two drawbacks as follows: One is that the IC unit requires divider units, which causes degradation of operating clock frequency. A computational cost of division is much higher than multiplication. A super-pipelined structure can recover clock frequency, however goes up its hardware size. The other is that the successive computation of (2) requires many computation cycles to generate all of $\tilde{x}_k$. For instance, the computation of $\tilde{x}_3$ in the third line in Fig. 4 starts after computing $\tilde{x}_1$ and $\tilde{x}_2$. It takes many pipeline latency cycles.

2) Proposed architecture: Note that the divisors and the dividends in the IC unit come from elements in the lower triangular matrix $\boldsymbol{L}$ in (1). We shift the divisions from the IC unit to the LQD unit as

$$\begin{aligned} L_{21} &= \frac{l_{21}}{l_{22}} \qquad L_{31} = \frac{l_{31}}{l_{33}} \qquad L_{32} = \frac{l_{32}}{l_{33}} \\ L_{41} &= \frac{l_{41}}{l_{44}} \qquad L_{42} = \frac{l_{42}}{l_{44}} \qquad L_{43} = \frac{l_{43}}{l_{44}}. \end{aligned} \tag{5}$$

Since the frequency of updating the CSI is low, the frequency of $\boldsymbol{L}$ is also low. The precomputation in (5) can be performed by the ASIP in the LQD unit. Next, we apply an arrayed structure in the IC unit by inserting waiting blocks illustrated in Fig. 7. The waiting blocks of "W1" enables exchanging data among parallel computing lines of $\tilde{x}_1$ to $\tilde{x}_4$. The intermediate result data of $\tilde{x}_1$, $\tilde{x}_2$, and $\tilde{x}_3$ can be used for the parallel computing lines. For the WCM unit, we make use of different timings in generating $\tilde{x}_1$ to $\tilde{x}_4$, whose structure is shown in Fig. 8. The data from $\tilde{x}_1$ to $\tilde{x}_4$ are generated for every two cycles. The proposed architecture inserts the waiting blocks of "W1" and simultaneously computes $\hat{x}_1$ to $\hat{x}_4$. It requires register units for delaying data, however can reduce pipeline latency cycles in the whole. The timing charts of the conventional and the proposed architectures are compared in Fig. 9. The proposed architecture reduces pipeline stages in the IC unit and enables concurrent processing in the IC and WCM units.

IV. Performance Evaluation

A. LQD unit

The TH precoder has been synthesized on 90 nm CMOS standard cell library where the supply voltage is 1.0 V. Table I shows circuit performance of the LQD unit. We set the clock frequency to 400 MHz. This evaluation includes not only LQ decomposition but also the precomputation of division employed in the IC unit for the proposed architecture. For the 160 MHz channel utilization in the IEEE802.11ac, the number of channel matrices (corresponding to OFDM data subcarriers) is 480. The total calculation time is 2.5 ns $\times$ 232.52 $\times$ 480 = 0.279 ms. This time is much shorter than the CSI update interval of 20 ms. This indicates that the LQD unit can provide real-time processing including the precomputation of division.

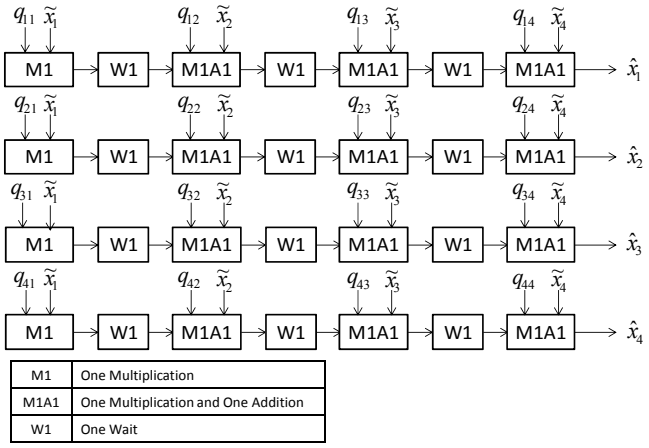


Fig. 8. Proposed architecture in WCM unit.

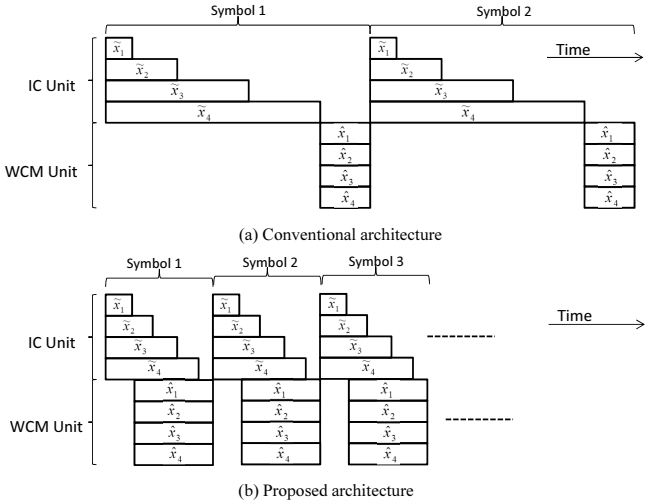


Fig. 9. Timing charts of conventional and proposed architectures.

B. IC and WCM units

Table II shows circuit performance of the conventional and the proposed architectures in the IC and the WCM units. The IC and WCM units have a 15-bit length in fixed-point arithmetic units. The target clock frequency is set to 160 MHz for all the units in logic synthesis and power consumption measurement. The gate-level power measured using a Synopsys Power Compiler in the condition of 1.0 V supply power. The proposed architecture exhibits smaller circuit area and power consumption because the conventional architecture needs many parallel structures of logic gates in logic synthesis to reduce a critical path delay.

V. Conclusion

We presented an arrayed pipelined TH precoder consisting of the LQD, the IC, and the WCM units for MU-MIMO systems. The LQD unit is designed by using an ASIP architecture. In the IC and the WCM units, the proposed architecture shortened a critical path and reduced circuit area and power consumption by 11% and 15%, respectively.

TABLE I
Performance of LQD unit.

Clock Frequency [MHz]	400
Gate Count	92,725
Power Consumption [mW]	39.4
Conputation Cycles per Matrix	232.52
Computation Time for CSI Interval [ms]	0.279
Energy Consumption [μJ]	11.0

TABLE II
Performance of IC and WCM units

	Conventional	Proposed
Maximum Clock Frequency [MHz]	160	160
Gate Count	144,518	128,368
Power Consumption [mW]	10.68	9.09
Computation Time for CSI Interval [ms]	10	10
Energy Consumption [μJ]	106.8	90.1

Acknowledgement

The authors would like to thank Prof. Shingo Yoshizawa, Kitami Institute of Technology, and the VLSI Design Education and Research Center (VDEC), Tokyo University for fruitful discussions. This study is supported in parts by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (A1) (24240007), the Japan Science and Technology Agency for A-Step Program (AS2416901H) and KDDI Laboratories.

References

[1] Q. H. Spencer, C. B. Peel, A. L. Swindlehurst, M. Haardt, "An introduction to the multi-user MIMO downlink," IEEE Communications Magazine, vol.42, no.10, pp.60-67, Oct. 2004.
[2] Y. S. Cho, J. Kim, W. Y. Yang, C. G. Kang, "MIMO-OFDM wireless communications with MATLAB," John Wiley & Sons (Asia) Pte Ltd, pp.408-417, 2010.
[3] C. B. Peel, B. M. Hochwald, A. L. Swindlehurst, "A vector- perturbation technique for near-capacity multiantenna multiuser communication- part I: channel inversion and regularization," IEEE Transactions on Communications, Vol.53, Issue 1, pp.195-202, Jan. 2005.
[4] C. Windpassinger, R. F. H. Fischer, T. Vencel, J. B. Huber, "Precoding in multiantenna and multiuser communications," IEEE Transactions on Wireless Communications, Vol.3, Issue 4, pp.1305-1316, July 2004.
[5] K. H. Lin, H. L. Lin, R. C. Chang, C. F. Wum, "Hardware architecture of improved Tomlinson-Harashima precoding for downlink MC-CDMA," IEEE Asia Pacific Conference on Circuits and Systems (APCCAS), pp.1200-1203, Dec. 2006.
[6] Y. Gu, K. K. Parhi, "High-speed architecture design of Tomlinson-Harashima precoders," IEEE Transaction on Circuits and Systems-I, Vol. 54, No. 9, pp. 1929-1937, Sep. 2007.
[7] Y. Hatakawa, T. Matsumoto, S. Konishi, "Development and experiment of linear and non-linear precoding on an real-time multiuser-MIMO testbed with limited CSI feedback," IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Sep. 2012.
[8] H. Iwaizumi, S. Yoshizawa, Y. Miyanaga, "A new high-speed and low-power LSI design of of SVD-MIMO-OFDM systems," IEEE International Symposium on Communications and Information Technologies (ISCIT), Oct. 2012.
[9] N.Shapira, Y. Shany, "Channel dimension reduction in MU operation," IEEE802.11 document 10/083r0, Jul. 2010.