PAPER Special Section on Smart Multimedia & Communication Systems

A VLSI Design of a Tomlinson-Harashima Precoder for MU-MIMO Systems Using Arrayed Pipelined Processing

Kosuke SHIMAZAKI^{†a)}, Student Member, Shingo YOSHIZAWA^{††}, Yasuyuki HATAKAWA^{†††}, Tomoko MATSUMOTO^{†††}, Satoshi KONISHI^{†††}, Members, and Yoshikazu MIYANAGA[†], Fellow

SUMMARY 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. In the IC and WCM units with fixed-point arithmetic, the proposed architecture uses an arrayed pipeline structure to shorten a circuit critical path delay. The implementation result shows that the proposed architecture reduces circuit area and power consumption by 11% and 15%, respectively.

key words: multi-user MIMO, Tomlinson-Harashima precoding, LQ decomposition, interference cancellation

1. 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 that is adopted in wireless LAN standard of IEEE 802.11n. The next-generation standard of IEEE 802.11ac is being formulated to achieve more than 1 Gbps throughputs and to support multi-user MIMO (MU-MIMO) to enhance communication capacity. MU-MIMO achieves 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 multiply 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

Manuscript received January 22, 2013.

Manuscript revised June 4, 2013.

a) E-mail: shimazaki@icn.ist.hokudai.ac.jp DOI: 10.1587/transfun.E96.A.2114 in linear precoding [3], [4]. The non-linear precoding has better communication quality but requires larger computational complexity.

Hardware architectures of THP have been presented by Lin et al. [5] and Gu and Parhi [6]. Their architectures focus on single-input and single-output (SISO) systems. THP testbed using a digital signal processor (DSP) for MU-MIMO has developed [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 application specific instruction-set processor (ASIP) architecture with floating-point arithmetic for high accuracy operations. We refer to the ASIP for singular value decomposition (SVD) designed in our previous work [8]. This ASIP has been developed for high-speed computation of SVD in SU-MIMO beamforming. In the IC and WCM units with fixed-point arithmetic, the conventional architecture takes a long circuit delay in successive interference cancellation of MIMO-THP. The proposed architecture uses an arrayed structure to have a shorter circuit delay 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 Sect. 2. Section 3 presents a design of the THP circuit. Section 4 describes performance evaluation of the designed circuit. Section 5 concludes the paper.

2. Tomlinson-Harashima Precoding for MU-MIMO Systems

A MU-MIMO system with four transmit antennas and two double-antenna users, i.e., 4×2 MU-MIMO, is illustrated in Fig. 1. We assume that channel state information (CSI) is ideally fed back from a receiver to a transmitter. In THP, signals are transmitted after subtraction of multi-user interference caused by the propagation channel. The channel matrix \boldsymbol{H} consisting of $\boldsymbol{H_1}$ and $\boldsymbol{H_2}$ is estimated in a receiver and decomposed into a lower triangular matrix \boldsymbol{L} and a unitary matrix \boldsymbol{O} .

[†]The authors are with the Graduate School of Information Science and Technology, Hokkaido University, Sapporo-shi, 060-0814 Japan.

^{††}The author is with the Department of Electrical and Electronic Engineering, Kitami Institute of Technology, Kitami-shi, 090-8507 Japan.

^{†††}The authors are with KDDI R&D Laboratories Inc., Fujimino-shi, 356-8502 Japan.

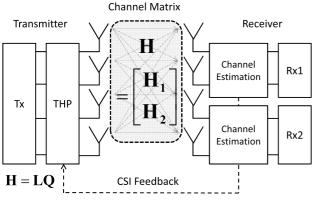


Fig. 1 4×2 MU-MIMO system.

$$H = LQ$$

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

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

$$\begin{cases}
\tilde{x}_{1} = Mod(x_{1}) = x_{1} \\
\tilde{x}_{2} = Mod(x_{2} - \frac{l_{21}}{l_{22}}\tilde{x}_{1}) \\
\tilde{x}_{3} = Mod(x_{3} - \frac{l_{31}}{l_{33}}\tilde{x}_{1} - \frac{l_{32}}{l_{33}}\tilde{x}_{2}) \\
\tilde{x}_{4} = Mod(x_{4} - \frac{l_{41}}{l_{44}}\tilde{x}_{1} - \frac{l_{42}}{l_{44}}\tilde{x}_{2} - \frac{l_{43}}{l_{44}}\tilde{x}_{3}).
\end{cases} (2)$$

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

$$\tilde{\mathbf{y}} = HW\tilde{\mathbf{x}} + \mathbf{n}$$

$$= LQQ^H\tilde{\mathbf{x}} + \mathbf{n}$$

$$= L\tilde{\mathbf{x}} + \mathbf{n},$$
(3)

where n is a white Gaussian noise. Here, modulo operation in (2) is denoted as

$$Mod(x) = x - floor\left(\frac{x + M + jM}{2M}\right) \times 2M,$$
 (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 worsens if the signals are increased by multiplying l_{ij} . A modulo window size depends on modulation levels [2]. The Gram-Schmidt orthonormalization algorithm is adopted for the LQ decomposition (LQD). In the algorithm, orthonormal vectors are calculated from given linear independent vectors. There are other methods for LQD, such as Givens rotation [9] and Householder transformation [10]. Compared with these methods, Gram-Schmidt orthonormalization is suitable for hardware implementation owing to the small calculation complexity [11].

In the Gram-Schmidt algorithm, we explain by an $n \times n$ square matrix A with n-dimensional row vectors given by

linear independent. A unitary matrix Q with row vectors q_n is generated by orthonormalizing each row vector a_n of A as

$$\boldsymbol{v}_{n} = \boldsymbol{a}_{n} - \sum_{k=1}^{n-1} (\boldsymbol{q}_{k}, \boldsymbol{a}_{n}) \, \boldsymbol{q}_{k}$$

$$\boldsymbol{q}_{n} = \frac{\boldsymbol{v}_{n}}{\|\boldsymbol{v}_{n}\|}.$$
(5)

A lower triangular matrix L is calculated by multiplying A by Q^H , which is generated in (5) as

$$AQ^{H} = \begin{bmatrix} a_{1} \\ a_{2} \\ \vdots \\ a_{n} \end{bmatrix} \begin{bmatrix} q_{1}^{H} & q_{2}^{H} & \cdots & q_{n}^{H} \end{bmatrix}$$

$$= \begin{bmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} \end{bmatrix} = L.$$
(6)

3. VLSI Design

3.1 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. 2. The LOD unit executes the Gram-Schmidt algorithm in Eqs. (5) and (6) with floating-point arithmetic operation because the matrix decomposition requires high accuracy and large dynamic range computations. The IC unit performs the inference cancellation in Eq. (2) and the WCM unit gives the matrix multiplication by $Q^H \tilde{x}$ in Eq. (3). The timing chart and the packet format in the TH precoder are illustrated in Fig. 3. They have very different requirements for calculation accuracy and throughput performance. The throughput requirement of the LQD unit is not high because the CSI does not change frequently. The CSI update interval of 20 ms is presented by Shapira and Shany [12]. On the other hand, the IC and WCM units request the same high throughput as the baseband symbol rate. Since the IEEE802.11ac standard supports 160-MHz channel utilization, the symbol rate of 160 symbols/seconds is required for real-time precoding, where its processing is denoted in the data symbols in Fig. 3.

To decrease a work load in the IC and WCM units, we apply pre-computation for lower triangular matrix L in Eq. (1). The divisions in Eq. (2) are shifted from the IC unit to the LQD unit as

$$L_{21} = \frac{l_{21}}{l_{22}} \qquad L_{31} = \frac{l_{31}}{l_{33}} \qquad L_{32} = \frac{l_{32}}{l_{33}}$$

$$L_{41} = \frac{l_{41}}{l_{44}} \qquad L_{42} = \frac{l_{42}}{l_{44}} \qquad L_{43} = \frac{l_{43}}{l_{44}}.$$
(7)

According to Eq. (7), Eq. (2) is rewritten as

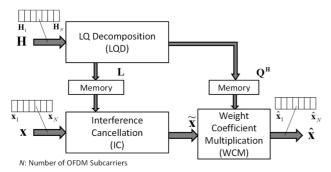


Fig. 2 Overall structure of TH precoder.

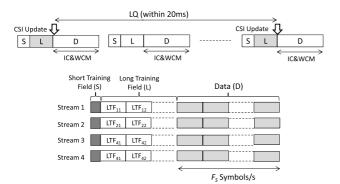


Fig. 3 Timing chart and packet format in TH precoder.

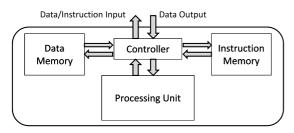


Fig. 4 Circuit structure of ASIP.

$$\begin{cases}
\tilde{x}_{1} = Mod(x_{1}) = x_{1} \\
\tilde{x}_{2} = Mod(x_{2} - L_{21}\tilde{x}_{1}) \\
\tilde{x}_{3} = Mod(x_{3} - L_{31}\tilde{x}_{1} - L_{32}\tilde{x}_{2}) \\
\tilde{x}_{4} = Mod(x_{4} - L_{41}\tilde{x}_{1} - L_{42}\tilde{x}_{2} - L_{43}\tilde{x}_{3}).
\end{cases}$$
(8)

Since the frequency of updating the CSI is low, the update frequency of \boldsymbol{L} is also low. The precomputation in Eq. (7) can be performed by the LQD unit.

3.2 ASIP Implementation of LQD Unit

The LQD unit is based on an application specific instruction-set processor (ASIP) architecture, which is illustrated in Fig. 4. We utilize the ASIP designed by Iwaizumi et al. [8], which provides high-speed computation of SVD in SU-MIMO beamforming. The same arithmetic operation units and instruction sets are effective in the LQ decomposition. The data and instructions are stored in each memory, and the processing unit executes instructions in order. Floating-point units (FPUs) deal with IEEE 754 standard

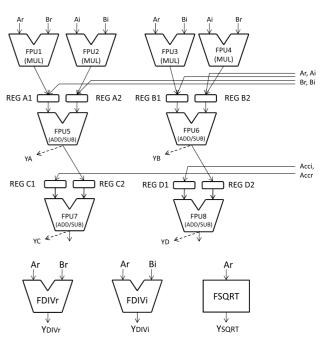


Fig. 5 Circuit structure of processing unit.

single precision floating-point in the processing unit. The circuit structure of the processing unit is illustrated in Fig. 5. The FPU supports four types of arithmetic operations: 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. Since all the FPUs execute pipeline processing, cycles per instruction (CPI) in this processor almost reaches to one by increasing block data size in pipeline processing. Table 1 enumerates the supported instructions. The instruction consists of memory address of input data A, B, output data C, and operation type, as illustrated in Fig. 6. The bit lengths are $\log_2 N$ and $\log_2 N_{op}$ bits, where N is the number of data memory words and N_{op} is the number of instructions. The dedicated highspeed division and square-root operation units denoted by "FDIV" and "FQRT" are also implemented. By reducing computation cycles, the total dissipated energy can be reduced at the cost of a smaller circuit area. Table 2 shows circuit performance of the LQD unit, where N=2,048 and N_{op} =256 are set in the processor specification. The LQD unit has been synthesized on a 90-nm CMOS standard cell library where the supply voltage is 1.0 V. 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 $2.5 \,\mathrm{ns} \times$ $232.52 \times 480 = 0.279$ ms. This time is much shorter than the CSI update interval of 20 ms. This indicates that the LO unit can provide real-time processing including the precom-

 Table 1
 Instructions supported in processing unit.

Table 1 mstractions supported in processing unit.			
OP	Instruction		
0	Complex addition		
1	Complex subtraction		
2	Complex multiplication		
3	Real multiplication		
4	Accumulative complex addition		
5	Accumulative complex subtraction		
6	Accumulative complex multiplication		
7	Accumulative real multiplication		
8	Real division		
9	Square-root operation		
10	Squared absolute value		
11	Accumulative squared absolute value		
12	Hermitian multiplication		
13	Initialization in Newton method		
14	Complex conjugate		
15	Data copy		
16	Initialization in CORDIC arctangent		
17	Repetition in CORDIC arctangent		
18	Initialization in CORDIC sine and cosine		
19	Repetition in CORDIC sine and cosine		
20	Merge real and imaginary parts		
21	Extraction of real part		
22	Extraction of imaginary part		
23	Extraction of sign parts		
24	Conversion from integer to floating-point format		
25	Conversion from floating-point format to integer		



Fig. 6 Instruction format.

 Table 2
 Performance of LQD unit.

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

putation of division.

3.3 Conventional Architecture

The conventional architecture is given by straightforward computation of Eq. (8). The structure of the conventional IC unit is shown in Fig. 7. The IC unit has operation blocks "M1S1" and "M2S2" consisting of multipliers and subtracters. The modulo block "Mod" performs arithmetic and floor operations in accordance with Eq. (4). The detailed structure of the modulo operation block is denoted in Fig. 8. The conventional architecture in the WCM units is illustrated in Fig. 9. The matrix operation of $Q^H \tilde{x}$ is done by complex multiplication and additions. The block of "A4" has four input ports and generates result data by one output port. The drawback is that the IC unit suffers from a long critical path to compute \tilde{x}_4 . The computation of \tilde{x}_4 requests the outputs of \tilde{x}_1 , \tilde{x}_2 , and \tilde{x}_3 . It decreases operating clock

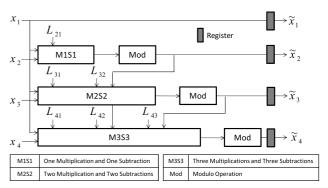


Fig. 7 Conventional architecture in IC unit.

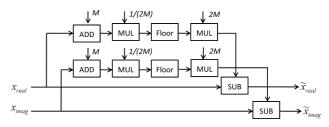


Fig. 8 Modulo operation block.

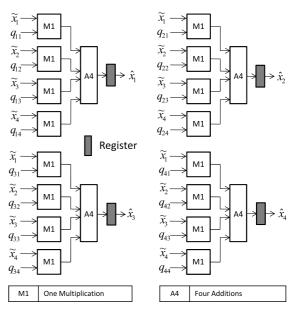


Fig. 9 Conventional architecture in WCM unit.

frequency or increases circuit area and power consumption by taking many parallel structures of logic gates in logic synthesis.

3.4 Proposed Architecture

The proposed IC unit with an arrayed structure is illustrated in Fig. 10. The three registers are inserted to shorten the critical path in the conventional IC unit. Due to the registers, the outputs of \tilde{x}_2 , \tilde{x}_3 , and \tilde{x}_4 are delayed in clock cycles. For the WCM unit, we use different timings to generate \tilde{x}_1 to

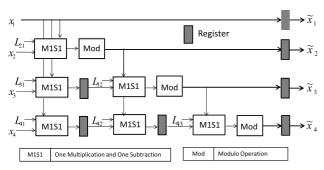


Fig. 10 Proposed architecture in IC unit.

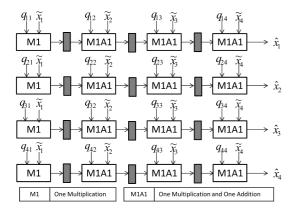
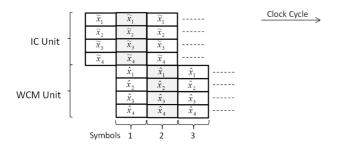
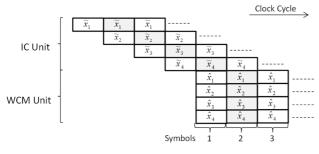


Fig. 11 Proposed architecture in WCM unit.



(a) Conventional architecture



(b) Proposed architecture

Fig. 12 Timing charts of conventional and proposed architectures.

 \tilde{x}_4 , the structure of which is shown in Fig. 11. The timing charts of the conventional and the proposed architectures are compared in Fig. 12. The proposed architecture takes the pipeline latency delay of three cycles unlike the conven-

Table 3 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

tional architecture. However, the throughput equal to the sampling rate does not change.

4. Evaluation

Table 3 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.

The power consumption of the LQD unit is larger than the summation of the IC and the WCM units. However, the LQD unit has much shorter computation time than the IC and the WCM units as shown in Table 2 and Table 3. We assume that the processing time of the IC and the WCM units in Fig. 3 occupies 50% for the CSI update interval. Hence, the LQD unit consumes much less energy than the IC and the WCM units.

5. 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. The proposed architecture in the IC and the WCM units shortened a critical path and reduced circuit area and power consumption by 11% and 15%, respectively.

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

 Q.H. Spencer, C.B. Peel, A.L. Swindlehurst, and M. Haardt, "An introduction to the multi-user MIMO downlink," IEEE Commun. Mag., vol.42, no.10, pp.60–67, Oct. 2004.

- [2] Y.S. Cho, J. Kim, W.Y. Yang, and 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, and A.L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication- part I: Channel inversion and regularization," IEEE Trans. Commun., vol.53, no.1, pp.195–202, Jan. 2005.
- [4] C. Windpassinger, R.F.H. Fischer, T. Vencel, and J.B. Huber, "Precoding in multiantenna and multiuser communications," IEEE Trans. Wireless Commun., vol.3, no.4, pp.1305–1316, July 2004.
- [5] K.H. Lin, H.L. Lin, R.C. Chang, and 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 and K.K. Parhi, "High-speed architecture design of Tomlinson-Harashima precoders," IEEE Trans. Circuits Syst. I, vol.54, no.9, pp.1929–1937, Sept. 2007.
- [7] Y. Hatakawa, T. Matsumoto, and 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), Sept. 2012.
- [8] H. Iwaizumi, S. Yoshizawa, and Y. Miyanaga, "A high-speed and low-energy-consumption processor for SVD-MIMO-OFDM systems," Hindawi, VLSI Design, vol.2013, Article ID 625019, March 2013.
- [9] S. Wang and E.E. Swartzlander, Jr., "The critically damped CORDIC algorithm for QR decomposition," IEEE Asilomar Conference on Signals, Systems and Computers, vol.2, pp.908–911, Nov. 1996.
- [10] S.-F. Hsiao and J.M. Delosme, "Householder CORDIC algorithms," IEEE Trans. Comput., vol.44, no.8, pp.990–1001, Aug. 1995.
- [11] R.C.-H. Chang, C.-H. Lin, K.-H. Lin, C.-L. Huang, and F.C.-Chen, "Iterative QR decomposition architecture using the modified Gram-Schmidt Algorithm for MIMO systems," IEEE Trans. Circuit Syst. I, vol.57, no.5, pp.1095–1102, May 2010.
- [12] N. Shapira and Y. Shany, "Channel dimension reduction in MU operation," IEEE802.11 document 10/083r0, July 2010.



Kosuke Shimazaki received a B.S. degree from Hokkaido University, Japan in 2012. He is currently studying at the Graduate School of Information Science and Technology, Hokkaido University. His research interests are wireless communication and VLSI design.



Shingo Yoshizawa received B.E., M.E., and Ph.D. degrees from Hokkaido University, Japan in 2001, 2003, and 2005, respectively. He was an assistant professor in the Graduate School of Information Science and Technology, Hokkaido University from 2006 to 2012. He is currently an associate professor in the Department of Electrical and Electronic Engineering, Kitami Institute of Technology. His research interests are speech processing, wireless communication, and VLSI architecture. He is a member

of IEEE and Research Institute of Signal Processing Japan.



Yasuyuki Hatakawa received B.S. and M.S. degrees from Hokkaido University, Hokkaido, Sapporo, Japan, in 2003 and 2005, respectively. He joined KDDI Corporation in 2005 and is currently involved in research and development of digital radio transmission techniques for mobile communications at KDDI R&D Laboratories. His current research interests include MIMO technology and digital signal processing for future communication systems.



Tomoko Matsumoto received the M.S. and Ph.D. degrees from the Yokohama National University, Japan, in 2005 and 2008, respectively. She joined KDDI Corporation in 2008. She is currently involved in research and development of digital radio transmission techniques for future mobile communications at KDDI R&D Laboratories.



Satoshi Konishi received B.S. and M.S. degrees in Electronic Engineering from the University of Electro-Communications (UEC), Tokyo, Japan, in 1991, and 1993, respectively. He also received a Ph.D. degree from Waseda University, Tokyo, Japan, in 2006. He joined Kokusai Denshin Denwa Co., Ltd. (now KDDI Corp.) in 1993. Since 1995, he has been engaged in research and development of radio resource allocation and management for wireless systems such as non-geostationary Earth orbit

mobile satellite systems, fixed wireless access systems and cellular systems. He is the senior manager in the Wireless Communications System Laboratory in KDDI R&D Laboratories Inc. His current research interests include the optimization of radio resources allocation, radio resource management, cross-layer control, multiple access techniques, adaptive transmission techniques, and adaptive signal processing for mobile communications systems. Dr. Konishi received the Young Researchers' Award from IEICE and the "Meritorious Award on Radio" from the Association of Radio Industries and Businesses (ARIB) in 2000 and 2010, respectively. He is a member of IEEE.



Yoshikazu Miyanaga is a professor in the Graduate School of Information Science and Technology, Hokkaido University. He is an associate editor of Journal of Signal Processing, RISP Japan (2005-present). He was a chair of the Technical Group on Smart Info-Media System, IEICE (IEICE TGSIS) (2004–2006) and is now a member of the advisory committee, IEICE TGSIS (2006-present). He is a vice-president of the Asia-Pacific Signal and Information Processing Association (APSIPA). He

was a distinguished lecture (DL) of IEEE CAS Society (2010–2011) and is now a member of the Board of Governors (BoG) of the same society (2011-present).