Iterative QR Decomposition Architecture Using the Modified Gram–Schmidt Algorithm for MIMO Systems

Robert Chen-Hao Chang, *Senior Member, IEEE*, Chih-Hung Lin, Kuang-Hao Lin, Chien-Lin Huang, and Feng-Chi Chen

Abstract—Implementation of an iterative QR decomposition (QRD) (IQRD) architecture based on the modified Gram-Schmidt (MGS) algorithm is proposed in this paper. A QRD is extensively adopted by the detection of multiple-input–multiple-output systems. In order to achieve computational efficiency with robust numerical stability, a triangular systolic array (TSA) for QRD of large-size matrices is presented. In addition, the TSA architecture can be modified into an iterative architecture that is called IQRD for reducing hardware cost. The IQRD hardware is constructed by the diagonal and the triangular process with fewer gate counts and lower power consumption than TSAQRD. For a 4×4 matrix, the hardware area of the proposed IQRD can reduce about 41% of the gate counts in TSAQRD. For a generic square matrix of order m IQRD, the latency required is $2m - 1$ time units, which is based on the MGS algorithm. Thus, the total clock latency is only $10m - 5$ cycles.

Index Terms— K-best detection, modified Gram–Schmidt (MGS), multiple-input–multiple-output (MIMO), orthogonal frequency-division multiplexing (OFDM), QR decomposition (QRD), triangular systolic array (TSA).

I. INTRODUCTION

A RECENT surge of research on wireless local area networks has given us new challenges and opportunities. With the increasing usage of wireless communication systems, reliability requirements for high data rates have become more critical. In the current decade, multiple-input-multiple-output (MIMO) systems have generated tremendous research interest as they offer high reliability and high throughput [1]–[5]. Therefore, MIMO has become a significant research for several wireless communication standards, for instance, the IEEE 802.11n and the IEEE 802.16e-2005, also referred to as mobile WiMAX. These wireless communication systems based on orthogonal frequency-division multiplexing (OFDM) will employ packet-based communication suitable for emerging broadband wired and wireless transmissions [6]–[8]. Next-generation wireless and mobile systems are currently the focus of research and development, which combine MIMO with OFDM. Consequently, MIMO–OFDM has become a popular technique [9]–[13].

To exploit the full potential of gains offered by MIMO, the computationally efficient design of a wireless baseband communication receiver has become difficult and challenging. A detection circuit involved in a MIMO receiver has to be designed for high data throughput [14]–[17]. The computational accuracy of the MIMO detection has a direct consequence on the throughput and reliability achieved in the receiver. Accordingly, QR decomposition (QRD) for a MIMO detection preprocessor is an essential component of all MIMO receivers [18]–[20]. The QRD processes the channel response $\mathbf{H}$ first and then decomposes it into $\mathbf{Q}$ and $\mathbf{R}$ matrices to produce $\mathbf{H} = \mathbf{QR}$.

Related studies on the QRD architecture can be classified into two major hardware implementation categories. The first category is the triangular systolic array (TSA) based on the Givens rotation algorithm and the coordinate rotation digital computer (CORDIC) algorithm [21], [22]. The Givens rotation algorithm can effectively reduce hardware area, but it has brought about longer clock latency in the QRD procedure. The other category is a parallel architecture based on the modified Gram-Schmidt (MGS) algorithm [23], [24]. However, within the extensive literature on QRD, relatively few studies focus on improving the clock latency and hardware area.

This paper proposes an iterative QRD (IQRD) hardware architecture based on the MGS algorithm. From the hardware point of view, the IQRD is constructed using the modified TSA architecture with lower gate count and clock latency than TSA structures. The IQRD architecture uses an iteration operation to reduce hardware complexity via diagonal-process (DP), triangular-process (TP), and dual-mode-process (DMP) circuits. The DMP circuit combines the DP with the TP and reduces unnecessary area in QRD hardware design. Simulation results show the performances of QRD and MIMO detection at different word-length numbers.

This paper is organized as follows. The MIMO detection is described in Section II. Section III discusses the QRD algorithms. In Section IV, we present the efficient QRD architecture design. In Section V, the simulation and implementation results are presented. Conclusions are drawn in Section VI.

R. C.-H. Chang and C.-H. Lin are with National Chung Hsing University, Taichung 402, Taiwan (e-mail: chchang@dragon.nchu.edu.tw).

K.-H. Lin is with the National Chin-Yi University of Technology, Taiping City 411, Taiwan (e-mail: khlin@ncut.edu.tw).

C.-L. Huang is with the National Chip Implementation Center, Hsinchu 300, Taiwan (e-mail: alex731114@gmail.com).

F.-C. Chen is with the Industrial Technology Research Institute, Hsinchu 31040, Taiwan (e-mail: chenfengchi@itri.org.tw).
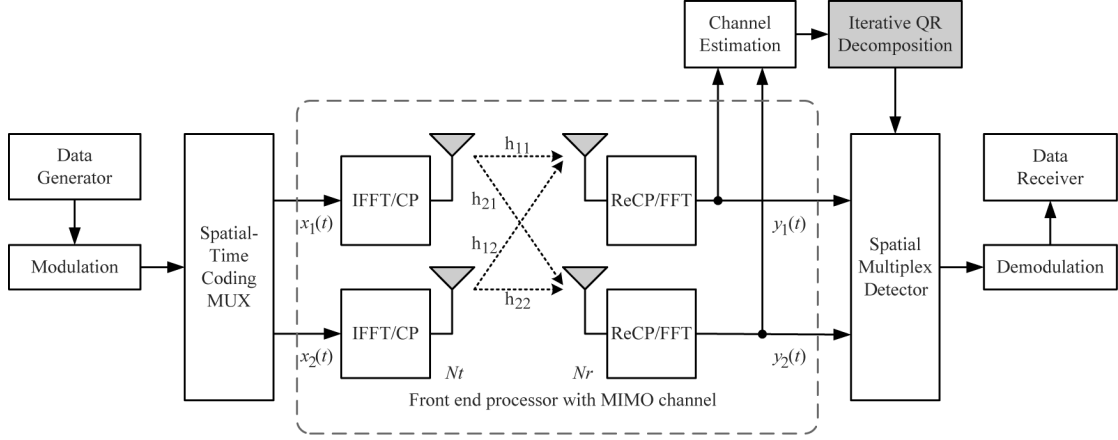
Fig. 1. Block diagram of a spatial multiplex MIMO wireless communication system.

II. MIMO DETECTION

A. MIMO System Model

To increase the data rate of wireless communication, the spatial multiplex is used in the MIMO system with Nt transmitted antennas and Nr received antennas. The block diagram of the MIMO system is shown in Fig. 1. The source data passes through modulation, spatial time coding, and inverse fast Fourier transform and cyclic prefix to transmit antennas in the transmitter. The receiver site includes fast Fourier transform and removed cyclic prefix, channel estimation, QRD, spatial multiplex detection, and demodulation to restore the source data. The baseband equivalent model can be described in

$$\mathbf{Y} = \mathbf{H}\mathbf{S} + \mathbf{n}. \tag{1}$$

At each symbol time, a vector $\mathbf{S} = [s_1, s_2, \ldots, s_{Nt}]^{\mathrm{T}}$ with each symbol belonging to the q quadrature amplitude modulation (q-QAM) constellation passes through the channel response $Nr \times Nt$ matrix $\mathbf{H}$. The received vector $\mathbf{Y} = [y_1, y_2, \ldots, y_{Nr}]^{\mathrm{T}}$ at the receiving antenna for each symbol time is a noisy superimposition of the Nt signals contaminated by additive white Gaussian noise (AWGN).

The complex matrix (1) can be transformed to its real matrix representation as

$$\begin{bmatrix} \Re(Y) \\ \Im(Y) \end{bmatrix} = \begin{bmatrix} \Re(H) & -\Im(H) \\ \Im(H) & \Re(H) \end{bmatrix} \begin{bmatrix} \Re(S) \\ \Im(S) \end{bmatrix} + \begin{bmatrix} \Re(n) \\ \Im(n) \end{bmatrix}. \tag{2}$$

B. K-Best Detector With QRD

We concentrate ourselves on the symmetric case, i.e., $Nr = Nt$, and use the QRD of $\mathbf{H}$. With $\mathbf{H} = \mathbf{Q}\mathbf{R}$, (1) can be rewritten as

$$\begin{aligned} \hat{\mathbf{y}} = \mathbf{Q}^{\mathrm{H}}\mathbf{Y} &= \mathbf{Q}^{\mathrm{H}}(\mathbf{H}\mathbf{S} + \mathbf{n}) \\ &= \mathbf{Q}^{\mathrm{H}}(\mathbf{Q}\mathbf{R}\mathbf{S} + \mathbf{n}) = \mathbf{R}\mathbf{S} + \mathbf{Q}^{\mathrm{H}}\mathbf{n} \end{aligned} \tag{3}$$

where $\mathbf{R}$ is the upper triangular matrix and $\mathbf{Q}$ is an $Nr \times Nt$ unitary matrix.

The maximum likelihood (ML) detector is the optimum detection algorithm for the MIMO system. It requires finding the signal point $\mathbf{S}$ from all transmit vector signal sets that minimize the Euclidean distance with respect to the received signal vector $\mathbf{y}$. With QRD, we can have

$$\hat{s} = \arg\min \|\mathbf{Y} - HS\|^2 = \arg\min \|\hat{\mathbf{y}} - RS\|^2. \tag{4}$$

Expanding the vector norm in (4) yields

$$\hat{s} = \arg\min \sum_{i=1}^{Nt} \left\| \hat{y}_i - R_{ii}S_i - \sum_{j=i+1}^{Nt} R_{ij}S_j \right\|^2. \tag{5}$$

The detection process starts from the last layers $l = Nt$ and works the way until the first layer is detected. It is to perform an exhaustive search of all possible combinations of the transmitted symbols that minimizes $\|\mathbf{Y} - \mathbf{H}\mathbf{S}\|^2$.

The branch cost function associated with nodes in the ith layer is

$$T_i(s^i) = T_{(i+1)}\left(s^{(i+1)}\right) + \left|e_i(s^i)\right|^2$$

$$\text{with} \quad e_i(s^i) = \hat{y}_i - R_{ii}S_i - \sum_{j=i+1}^{Nt} R_{ij}S_j. \tag{6}$$

Each node in the tree corresponds to a so-called partial Euclidean distance (PED) $T_i(s^{(i)})$, where $T_{Nt+1}(s^{(Nt+1)}) = 0$ and the term $|e_i(s^{(i)})|^2$ denotes the distance increment between two successive nodes in the tree. At each layer, the K-best detector approximates a breadth-first search by keeping only K candidates with the smallest PEDs for the next level search.

III. QRD

The common QRD algorithms are Givens rotation [25], MGS orthogonalization [26], Householder transformation [27], etc. Since Householder transformation is much more complex in hardware implementation than the other two algorithms, this paper only discusses the Givens rotation and the MGS orthogonalization.

A. Givens Rotation

The Givens rotation rotates in the plane expanded by two co-ordinate axes. The rotation matrix can be expressed as

$$
\mathbf{G}(i,k,\theta) = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & \cos\theta_{i,i} & \cdots & -\sin\theta_{i,k} & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & \sin\theta_{k,i} & \cdots & \cos\theta_{k,k} & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix}. \tag{7}
$$

Here, $\mathbf{G}(i,k,\theta)^{\mathrm{T}} \cdot X$ denotes that the vector X counterclockwise rotates an angle θ in the (i,k) plane. When the rotation matrix $\mathbf{G}$ multiplies another matrix $\mathbf{A}$, it only affects the ith and kth row of matrix $\mathbf{A}$.

Given a nonsingular $n \times n$ matrix $\mathbf{A}$, we use Givens rotation to obtain the QR factorization of $\mathbf{A}$. Let $\mathbf{G}_{21}$ be the Givens rotation acting on the first and second coordinates so that a zero results in the (2, 1) position. We can apply another Givens rotation $\mathbf{G}_{31}$ to $(\mathbf{G}_{21} \bullet \mathbf{A})$ to obtain a zero in the (3, 1) position. The process can be continued until the last $n - 1$ entry in the first column has been eliminated

$$
\mathbf{G}_{n1} \cdots \mathbf{G}_{31} \mathbf{G}_{21} \mathbf{A} = \begin{bmatrix} \times & \times & \cdots & \times \\ 0 & \times & \cdots & \times \\ 0 & \times & \cdots & \times \\ \vdots & & & \\ 0 & \times & \cdots & \times \end{bmatrix}. \tag{8}
$$

Next, Givens rotations $\mathbf{G}_{32}, \mathbf{G}_{42}, \ldots, \mathbf{G}_{n2}$ are used to eliminate the last $n - 2$ entries in the second column. This process is continued until all elements below the diagonal have been eliminated

$$
(\mathbf{G}_{n,n1}) \cdots (\mathbf{G}_{n2} \cdots \mathbf{G}_{32})(\mathbf{G}_{n1} \cdots \mathbf{G}_{21})\mathbf{A} = \mathbf{R} \tag{9}
$$

where $\mathbf{R}$ is an upper triangular matrix. If we let $\mathbf{Q}^{\mathrm{T}} = (\mathbf{G}_{n,n-1}) \cdots (\mathbf{G}_{n2} \cdots \mathbf{G}_{32})(\mathbf{G}_{n1} \cdots \mathbf{G}_{21})$, then $\mathbf{A} = \mathbf{QR}$ and $\mathbf{A}\boldsymbol{x} = \mathbf{b}$ is equivalent to $\mathbf{R}\boldsymbol{x} = \mathbf{Q}^{\mathrm{T}}\mathbf{b}$.

Based on (7), if QRD is to be executed on any matrix, an upper triangular matrix $\mathbf{R}$ can be obtained only through successive multiplication of the rotation matrix $\mathbf{G}$. The orthogonal matrix $\mathbf{Q}$ can also be obtained by executing the same operation on identity matrix $\mathbf{I}$. When the QRD is achieved by the Givens rotation algorithm, the hardware complexity can be reduced by the CORDIC operation core with vectoring and rotating modes.

QRD can be realized by using the CORDIC vectoring mode to convert the complex input matrix numbers into real numbers and do row cancellation. This produces the matrix $\mathbf{R}$, all of whose diagonal elements are real numbers. Mathematically, assuming that the input is equal to $x + iy$, the CORDIC vectoring mode is utilized to eliminate y. In other words, use each calculated value of y to control the direction of rotation. The rotating mode makes a coordinate rotate a given angle or two coordinates rotate the same angle. For QRD, the rotating angle θ produced by the vectoring mode is transmitted to the other elements of the same row for coordinate rotation. The direction of rotation can be controlled for every calculated value of θ.

The QRD architecture using the Givens rotation algorithm can effectively reduce hardware area, because this algorithm employs an iterative architecture to implement the vectoring- and rotating-mode circuits with the CORDIC. Therefore, the MIMO detection system needs high-speed components to perform QRD. In addition, using the Givens rotation algorithm has brought about longer clock latency in the QRD procedure.

B. MGS

The QRD for any matrix can be realized by the MGS orthogonalization algorithm [23]. This procedure reduces the memory hardware consumption. The MGS orthogonalization is described as follows.

a_{ij} denotes row i and column j of the matrix $\mathbf{A}$, and $\mathbf{a}_j$ denotes the column vector of the matrix $\mathbf{A}$. The QRD of the matrix $\mathbf{A}$, $\mathbf{A} = \mathbf{QR}$, can be obtained by the following steps. For simplicity, a 4×4 matrix $\mathbf{A}$ is utilized to explain the procedure. First, the first column vector $\mathbf{a}_1$ is normalized to obtain

$$
r_{11} = \|\mathbf{a}_1\|_2 = \sqrt{(a_{11})^2 + (a_{21})^2 + (a_{31})^2 + (a_{41})^2} \tag{10}
$$

which represents the element in row 1 and column 1 of $\mathbf{R}$. The value of $\mathbf{q}_1$, which is the first column vector of $\mathbf{Q}$, can be calculated from r_{11}

$$
\mathbf{q}_1 = \frac{\mathbf{a}_1}{r_{11}} \rightarrow q_{11} = \frac{a_{11}}{r_{11}}, q_{21} = \frac{a_{21}}{r_{11}}, q_{31} = \frac{a_{31}}{r_{11}}, q_{41} = \frac{a_{41}}{r_{11}}. \tag{11}
$$

Second, r_{12}, r_{13}, and r_{14} can be calculated using the first column vector of the matrix $\mathbf{Q}(\mathbf{q}_1)$ and the second, third, and fourth column vectors of the matrix $\mathbf{A}$

$$
\begin{aligned}
r_{1j} &= \mathbf{q}_1^{\mathrm{T}} \mathbf{a}_j \\
&= q_{11}a_{1j} + q_{21}a_{2j} + q_{31}a_{3j} + q_{41}a_{4j}, \\
& \qquad\qquad\qquad\qquad j \in N; 2 \leq j \leq 4. \tag{12}
\end{aligned}
$$

After obtaining $\mathbf{q}_1$, r_{12}, r_{13}, and r_{14} from (11) and (12), the matrix $\mathbf{A}$ can be converted to the matrix $\boldsymbol{A}^1 = [\mathbf{a}_1^1 \ \mathbf{a}_2^1 \ \mathbf{a}_3^1 \ \mathbf{a}_4^1]$

$$
\begin{aligned}
\mathbf{a}_1^1 &= 0 \\
\mathbf{a}_2^1 &= \mathbf{a}_2 - r_{12}\mathbf{q}_1 \\
\mathbf{a}_3^1 &= \mathbf{a}_3 - r_{13}\mathbf{q}_1 \\
\mathbf{a}_4^1 &= \mathbf{a}_4 - r_{14}\mathbf{q}_1. \tag{13}
\end{aligned}
$$

So far, these steps have derived the values of $\mathbf{q}_1$, r_{11}, r_{12}, r_{13}, r_{14}, and the matrix $\boldsymbol{A}^1$. The aforementioned steps are then repeated with $\boldsymbol{A}^1$ to compute $\mathbf{q}_2$, r_{22}, r_{23}, r_{24}, and the matrix $\boldsymbol{A}^2$. The second column vector of the matrix $\boldsymbol{A}^1$ is normalized to obtain r_{22}, i.e.,

$$
r_{22} = \|\mathbf{a}_2^1\|_2 = \sqrt{\left(a_{12}^1\right)^2 + \left(a_{22}^1\right)^2 + \left(a_{32}^1\right)^2 + \left(a_{42}^1\right)^2}. \tag{14}
$$

Then, $\mathbf{q}_2$, r_{23}, and r_{24} can be calculated

$$
\mathbf{q}_2 = \frac{\mathbf{a}_2^1}{r_{22}} \tag{15}
$$

$$
r_{2j} = \mathbf{q}_2^{\mathrm{T}} \mathbf{a}_j^1, \qquad j \in N; 3 \leq j \leq 4. \tag{16}
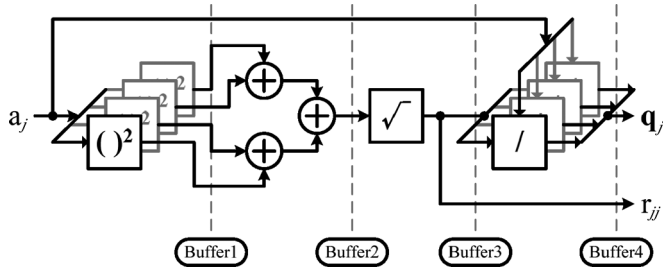$$

Fig. 2. Hardware architecture of the DP for QRD.

Fig. 3. Hardware architecture of the TP for QRD.

Fig. 4. DMP combines DP with TP.

The matrix $\mathbf{A}^2$ is obtained from the following:

$$\begin{aligned}
\mathbf{a}_1^2 &= 0 \\
\mathbf{a}_2^2 &= 0 \\
\mathbf{a}_3^2 &= \mathbf{a}_3^1 - r_{23}\mathbf{q}_2 \\
\mathbf{a}_4^2 &= \mathbf{a}_4^1 - r_{24}\mathbf{q}_2.
\end{aligned} \tag{17}$$

After obtaining $\mathbf{q}_2$, r_{22}, r_{23}, r_{24}, and the matrix $\mathbf{A}^2$, the matrix $\mathbf{A}^2$ can be used to obtain $\mathbf{q}_3$, r_{33}, r_{34}, and the matrix $\mathbf{A}^3$. The third column vector of the matrix $\mathbf{A}^2$ is normalized to get r_{33}, i.e.,

$$r_{33} = \left\| \mathbf{a}_3^2 \right\|_2 = \sqrt{\left(a_{13}^2\right)^2 + \left(a_{23}^2\right)^2 + \left(a_{33}^2\right)^2 + \left(a_{43}^2\right)^2}. \tag{18}$$

Then, $\mathbf{q}_3$ and r_{34} can be calculated

$$\mathbf{q}_3 = \frac{\mathbf{a}_3^2}{r_{33}} \tag{19}$$

$$r_{3j} = \mathbf{q}_3^{\mathrm{T}} \mathbf{a}_j^2, \qquad j = 4. \tag{20}$$

The matrix $\mathbf{A}^3$ is obtained from the following:

$$\begin{aligned}
\mathbf{a}_1^3 &= 0 \\
\mathbf{a}_2^3 &= 0 \\
\mathbf{a}_3^3 &= 0 \\
\mathbf{a}_4^3 &= \mathbf{a}_4^2 - r_{34}\mathbf{q}_3.
\end{aligned} \tag{21}$$

Similarly, by repeating these steps in the matrix $\mathbf{A}^3$, r_{44} can be computed by normalizing the fourth column vector of the matrix $\mathbf{A}^3$, i.e.,

$$r_{44} = \left\| \mathbf{a}_4^3 \right\|_2 = \sqrt{\left(a_{14}^3\right)^2 + \left(a_{24}^3\right)^2 + \left(a_{34}^3\right)^2 + \left(a_{44}^3\right)^2}. \tag{22}$$

Therefore

$$\mathbf{q}_4 = \frac{\mathbf{a}_4^3}{r_{44}}. \tag{23}$$

Finally, $\mathbf{A} = \mathbf{QR}$ is obtained, where

$$\mathbf{Q} = \begin{bmatrix} q_{11} & q_{12} & q_{13} & q_{14} \\ q_{21} & q_{22} & q_{23} & q_{24} \\ q_{31} & q_{32} & q_{33} & q_{34} \\ q_{41} & q_{42} & q_{43} & q_{44} \end{bmatrix} \quad \mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & r_{14} \\ 0 & r_{22} & r_{23} & r_{24} \\ 0 & 0 & r_{33} & r_{34} \\ 0 & 0 & 0 & r_{44} \end{bmatrix}.$$

IV. Architecture Design

The aforementioned derivation shows that the QRD can be separated into two steps. The first step computes the diagonal line elements of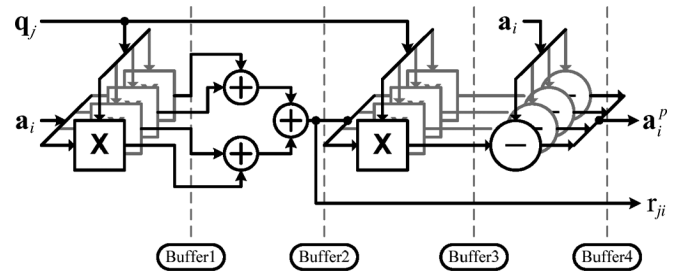 the upper triangular matrix $\mathbf{R}$ a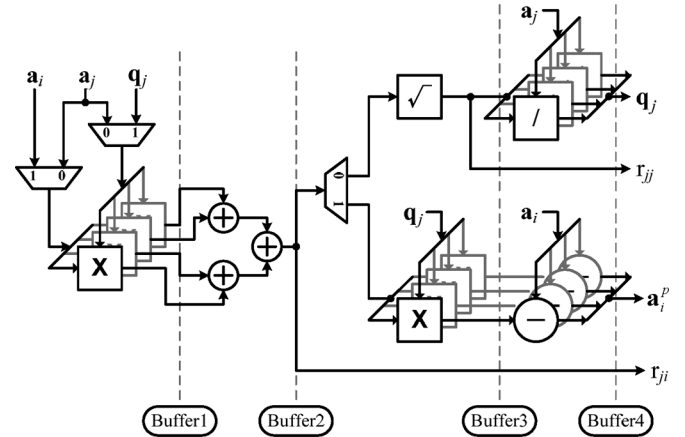nd the unitary matrix $\mathbf{Q}$. A diagonal line element (r_{jj}) indicates the square root of the summation of the square of each element in the column j of $\mathbf{a}(\mathbf{a}_j)$. To save hardware, $\mathbf{a}_j$ elements are imported to the DP in parallel. Four squarers, three adders, and a square-root operator are adopted for a 4×4 matrix, as shown in Fig. 2. Additionally, $\mathbf{a}_j$ elements are delayed by buffer registers and sequentially divided by r_{jj}. Therefore, $\mathbf{q}_j$ of the unitary matrix $\mathbf{Q}$ can be derived. To increase operation frequency, a four-stage pipeline structure can be constructed by using buffers.

In the second step, the $\mathbf{q}_j$ of DP is input to the TP in parallel, and the matrix $\mathbf{R}$ of the nondiagonal line elements (r_{ji}) with the new matrix $\mathbf{A}(\mathbf{a}_i^p)$ is computed. Fig. 3 shows the hardware architecture of TP for a 4×4 matrix, which uses eight multipliers, three adders, and four subtractors. DMP is proposed to combine DP with TP, and the hardware area is reduced by eliminating four multipliers and three adders, as shown in Fig. 4. If the select signals of the multiplexers and demultiplexer in DMP are "0," DP is active to compute r_{jj} and $\mathbf{a}_j$. If the select signals are "1," TP is active.

All of the element values in the matrices $\mathbf{Q}$ and $\mathbf{R}$ can be derived by the repetitive operations of DP and TP. In the following subsection, the TSAQRD for a 4×4 matrix will be presented. However, the hardware area of TSAQRD substantially increases as the rank of the matrix increases. Therefore, the IQRD using the feedback control circuit and hardware sharing to significantly reduce the hardware area will be proposed.

A. TSA

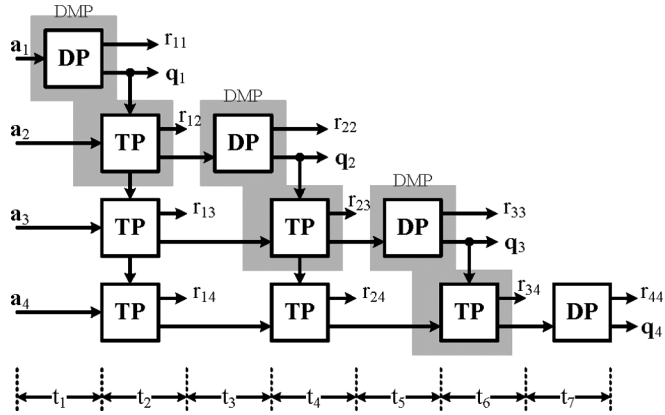Fig. 5 shows the TSA architecture for QRD with the MGS algorithm. The TSA consists of a set of processing units. Each

Fig. 5. TSAQRD for a 4 × 4 matrix.

Fig. 6. IQRD for a 4 × 4 matrix.

Fig. 7. Simulation result of the K-best algorithm with 64 QAM in the uncoded MIMO system ($Nr = Nt = 4$).

processing unit can perform some simple operations. The advantage of this architecture is a simple and regular design that can speed up computation flow. Using the TSA architecture for QRD hardware can offer high throughput, but the processing units increase with the dimension of $\mathbf{H}$. The TSAQRD processing units demand m DPs and $\sum_{i=1}^{m-1}(m-i)$ TPs if the dimension of the matrix $\mathbf{H}$ is $m \times m$.

The TSAQRD sequential operation needs seven time slots from t_1 to t_7 when a 4 × 4 matrix is decomposed. DP executes in odd time slots (t_1, t_3, t_5, t_7) and TP executes in even time slots (t_2, t_4, t_6). In the first time slot t_1, a_1 operates through the DP circuit, obtaining r_{11} and $\mathbf{q}_1$. $\mathbf{q}_1$ distributes through TP circuits. In addition, (a_2, a_3, a_4) passes through the delay buffer (B) and waits for $\mathbf{q}_1$ to operate in the TP circuit. In the next time slot t_2, $(a_2, \ldots, a_8)$ and $\mathbf{q}_1$ are computed by TP, and then, (r_{12}, r_{13}, r_{14}) and $(\mathbf{a}_2^1, \mathbf{a}_3^1, \mathbf{a}_4^1)$ are generated, respectively. In the time slot t_3, $\mathbf{a}_2^1$ in DP and $(\mathbf{a}_3^1, \mathbf{a}_4^1)$ sequentially feed back to the delay buffer. Therefore, repetitive operations using DP and TP accomplish QRD. The hardware area is defined as

$$A_{\text{TSAQRD}} = G_{\text{DP}}m + G_{\text{TP}}\sum_{i=1}^{m-1}(m-i) \qquad (24)$$

where G_{DP} is the gate count of a DP circuit and G_{TP} is the gate count of a TP circuit.

The hardware area of TSAQRD can be reduced by using DMP, which is designed by the hardware sharing of DP and TP. Using the DMP architecture can reduce $m - 1$ redundant circuits that include four multipliers and three adders. The hardware area A_{TSAQRD} can be modified as

$$A_{\text{TSAQRD}} = G_{\text{DMP}}(m-1) + G_{\text{DP}} + G_{\text{TP}}\sum_{i=1}^{m-2}(i) \qquad (25)$$

where G_{DMP} is the gate count of a DMP circuit.

B. IQRD

Fig. 6 shows the proposed IQRD architecture, which is based on TSAQRD and utilizes the feedback control circuit and the hardware sharing technique to reduce complexity and hardware area. The IQRD sequential operation also needs seven time slots. For a generic square matrix of order m, the DP latency
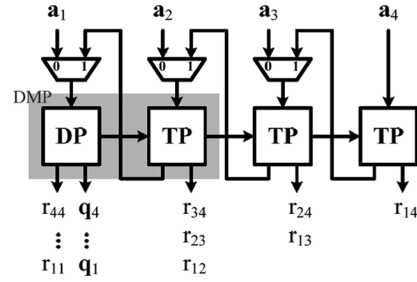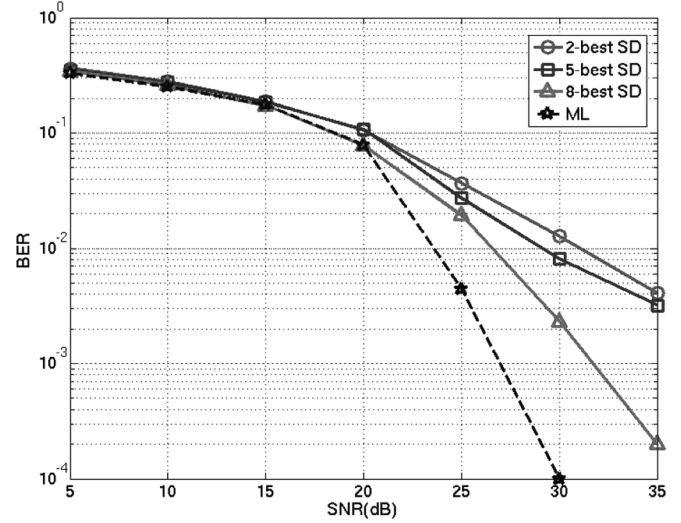
required is m time units, and the TP latency required is $(m-1)$ time units. Thus, the total time units (T_{tu}) for IQRD is defined as

$$T_{\text{tu}} = m + (m-1) = (2m-1). \qquad (26)$$

The IQRD architecture requires few latency clocks than the architecture proposed by Singh *et al.* [23]. The clock latency is five clocks for DP, TP, and DMP circuits. Therefore, the total number of the latency clocks (T_{lc}) is defined as

$$T_{\text{lc}} = 5T_{\text{tu}}. \qquad (27)$$

The hardware area of IQRD, i.e., A_{IQRD}, can be calculated as

$$A_{\text{IQRD}} = G_{\text{DMP}} + G_{\text{TP}}(m-2). \qquad (28)$$

Compared with the TSAQRD, the IQRD hardware can decrease $G_{\text{DMP}}(m-2) + G_{\text{DP}} + G_{\text{TP}}\sum_{i=1}^{m-3}(i)$ gate counts. The hardware area of the proposed IQRD reduces about 41% of the gate counts in TSAQRD for a 4 × 4 matrix.

V. SIMULATION AND IMPLEMENTATION RESULT

Fig. 7 compares the K-best detector with the ML detector while using the preprocessor of IQRD at the receiver of MIMO systems. Our simulation is over the AWGN Rayleigh independent and identically distributed channel model with
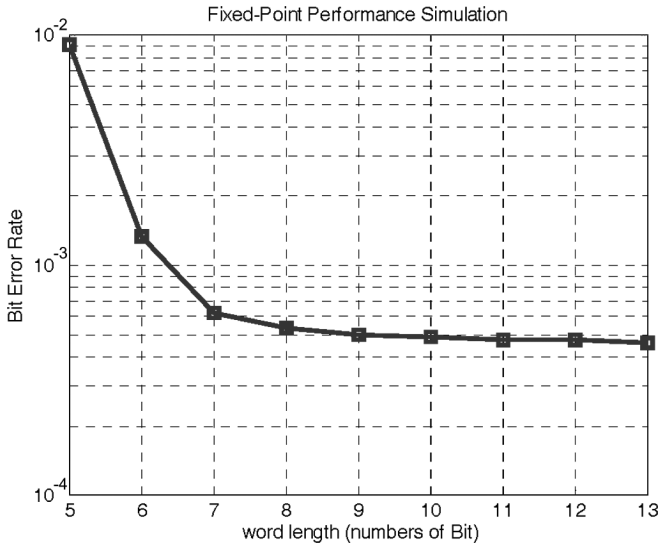
Fig. 8. Word-length simulation result of the K-best algorithm in the uncoded MIMO system ($Nr = Nt = 4$ and 64 QAM).
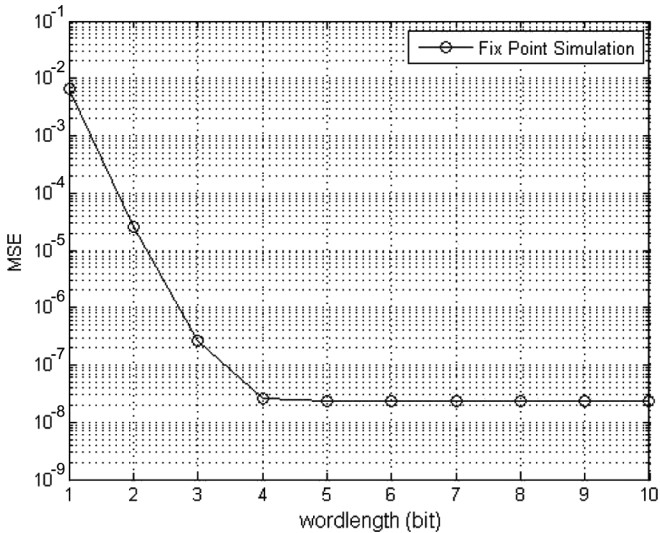

Fig. 9. Fixed-point simulation of IQRD with different word lengths.

multiantennas $Nr = Nt = 4$ and without channel coding. The fixed-point number is a finite word-length representation of the corresponding floating-point number. The fixed-point number is in the two's complement format, which includes one sign bit, four integer bits, and the fractional bits. Fig. 8 shows the analysis of the fractional word length for hardware design. If the fractional word length is longer than 7 bits, then the BER is saturated with an SNR of 33 dB. Therefore, fractional word lengths of 9 bits should be chosen for hardware implementation. To deal with the overflow issue in the divider, this architecture adopts 18 bits for fractional operation and then truncates the word lengths into 9 bits before output.

To determine the best number of bits for hardware implementation of IQRD, fixed-point simulation is also performed. Fig. 9 shows the mean square error (mse) versus fractional word length. The mse represents the error between the fixed- and floating-point QRD outputs. In Fig. 9, the curve is saturated with 10^{-8} mse. Increasing the number of bits does not improve

TABLE I
PERFORMANCE COMPARISON OF DP, TP, AND DMP

	Gate Count	Critical Path	Power consumption
DP	21.9K	2.2ns	5.9mW
TP	9.1K	2.2ns	4.8mW
DMP	26.1K	2.2ns	5.9mW

TABLE II
IMPLEMENTATION DIFFERENCES OF TSAQRD AND IQRD

	TSAQRD	IQRD
DP	1	-
TP	$\sum_{i=1}^{m-2}(i)$	m-2
DMP	m-1	1

TABLE III
COMPARISONS OF THE QRD IMPLEMENTATION RESULTS

	This work		Liu et al.[22]	Singh et al.[23]
Algorithm	MGS		Givens Rotation	MGS
	TSAQRD	IQRD		
Order	4×4	4×4	4×4	4×4
Structure	Feed forward	Iterative	Iterative	Feed forward
Clock Latency	35	35	264 (4×(4N+10))	67
Critical Path	2.5ns	2.5ns	3.3ns	4.4ns
Word length	14		20	16
Throughput	2.56Gbps		360Mbps	1.2Gbps
Technology	0.18-μm CMOS	0.18-μm CMOS	0.13-μm CMOS	0.18-μm CMOS
Gate Count	55.3K	32.6K	about 70K (4PE×17.5k)	72K
Memory	None	None	About 12K (4PE×3K)	None

system performance. Therefore, a 9-bit fractional word length is sufficient for the hardware design of IQRD.

The performance comparison of DP, TP, and DMP is listed in Table I. The DMP combines the DP with the TP to reduce about 15.8% gate counts. Table II shows the numbers of subblock implementation of the TSAQRD and the IQRD. Table III illustrates the comparison results of our works with Liu et al. [22] and Singh et al. [23]. Although the work of Liu et al. using the Givens rotation algorithm has similar gate count to the proposed IQRD structure, the memory required is 12 000 and more clock latency is needed. As Table III shows, the IQRD architecture has lower clock latency, higher throughput (2.56 Gbits/s), and smaller hardware area (gate count). Comparing the TSAQRD with the IQRD, the latter reduces gate count by 41% for a 4 × 4 matrix. More gate count can be saved for a larger matrix.

VI. CONCLUSION

The hardware architecture design of QRD is extensively discussed in current MIMO detection system studies on enhancing operational efficiency. The most popular architecture adopted is the Givens rotation algorithm with processing elements based on CORDIC computing. The Givens rotation algorithm

can effectively reduce hardware area, but it has brought about longer clock latency in the QRD procedure. In this paper, we have adopted the TSA structure to implement QRD with MGS, and then, it is modified by an iterative structure to achieve smaller clock latency and chip area. The total number of the latency clocks of IQRD is ($10m - 5$) cycles, and the hardware area is $G_{\mathrm{DMP}} + G_{\mathrm{TP}}(m - 2)$ gate counts. Compared with the TSAQRD, the IQRD hardware can decrease $G_{\mathrm{DMP}}(m - 2) + G_{\mathrm{DP}} + G_{\mathrm{TP}} \sum_{i=1}^{m-3}(i)$ gate counts. The hardware area of the proposed IQRD reduces about 41% of the gate counts in TSAQRD for a 4×4 matrix. The proposed architecture is implemented and verified by Taiwan Semiconductor Manufacturing Company 0.18-μm CMOS technology.

Acknowledgment

References

[1] A. J. Paulraj, D. A. Gore, R. U. Nabar, and H. Bolcskei, "An overview of MIMO communications—A key to gigabit wireless," *Proc. IEEE*, vol. 92, no. 2, pp. 198–218, Feb. 2004.

[2] H.-L. Lin, R. C. Chang, K.-H. Lin, and C.-C. Hsu, "Implementation of synchronization for 2×2 MIMO WLAN system," *IEEE Trans. Consum. Electron.*, vol. 52, no. 3, pp. 766–773, Aug. 2006.

[3] Q. Ling and T. Li, "Blind-channel estimation for MIMO systems with structured transmit delay scheme," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 55, no. 8, pp. 2344–2355, Sep. 2008.

[4] S. P. Alex and L. M. A. Jalloul, "Performance evaluation of MIMO in IEEE802.16e/WiMAX," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 2, pp. 181–190, Apr. 2008.

[5] J.-L. Yu and Y.-C. Lin, "Space-time-coded MIMO ZP-OFDM systems: Semiblind channel estimation and equalization," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 56, no. 7, pp. 1360–1372, Jul. 2009.

[6] L. Cimini, Jr., "Analysis and simulation of a digital mobile channel using orthogonal frequency division multiplexing," *IEEE Trans. Commun.*, vol. COM-33, no. 7, pp. 665–675, Jul. 1985.

[7] A. Troya, K. Maharatna, M. Krstic, E. Grass, U. Jagdhold, and R. Kraemer, "Low-power VLSI implementation of the inner receiver for OFDM-based WLAN systems," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 55, no. 2, pp. 672–686, Mar. 2008.

[8] K.-H. Lin, R. C. Chang, H.-L. Lin, and C.-F. Wu, "Analysis and architecture design of a downlink M-modification MC-CDMA system using the Tomlinson–Harashima precoding technique," *IEEE Trans. Veh. Technol.*, vol. 57, no. 3, pp. 1387–1397, May 2008.

[9] G. L. Stuber, J. R. Barry, S. W. McLaughlin, Y. Li, M. A. Ingram, and T. G. Pratt, "Broadband MIMO–OFDM wireless communications," *Proc. IEEE*, vol. 92, no. 2, pp. 271–294, Feb. 2004.

[10] H. Kim, J. Kim, S. Yang, M. Hong, and Y. Shin, "An effective MIMO–OFDM system for IEEE 802.22 WRAN channels," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 55, no. 8, pp. 821–825, Aug. 2008.

[11] H.-L. Lin, R. C. Chang, and H.-L. Chen, "A high speed SDM-MIMO decoder using efficient candidate searching for wireless communication," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 55, no. 3, pp. 289–293, Mar. 2008.

[12] L. Boher, R. Rabineau, and M. Helard, "FPGA implementation of an iterative receiver for MIMO–OFDM systems," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 6, pp. 857–866, Aug. 2008.

[13] M.-S. Baek, Y.-H. You, and H.-K. Song, "Combined QRD-M and DFE detection technique for simple and efficient signal detection in MIMO–OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 8, no. 4, pp. 1632–1638, Apr. 2009.

[14] C.-H. Yang and D. Markovic, "A flexible DSP architecture for MIMO sphere decoding," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 56, no. 10, pp. 2301–2314, Oct. 2009.

[15] C. S. Park, K. K. Parhi, and S.-C. Park, "Probabilistic spherical detection and VLSI implementation for multiple-antenna systems," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 56, no. 3, pp. 685–698, Mar. 2009.

[16] Z. Guo and P. Nilsson, "Algorithm and implementation of the K-best sphere decoding for MIMO detection," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 491–503, Mar. 2006.

[17] S. Chen, T. Zhang, and Y. Xin, "Relaxed K-best MIMO signal detector design and VLSI implementation," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 15, no. 3, pp. 328–337, Mar. 2007.

[18] J. Liu, Y. V. Zakharov, and B. Weaver, "Architecture and FPGA design of dichotomous coordinate descent algorithms," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 56, no. 11, pp. 2425–2438, Nov. 2009.

[19] C.-J. Ahn, "Parallel detection algorithm using multiple QR decompositions with permuted channel matrix for SDM/OFDM," *IEEE Trans. Veh. Technol.*, vol. 57, no. 4, pp. 2578–2582, Jul. 2008.

[20] T. H. Im, I. Park, J. Kim, J. Yi, J. Kim, S. Yu, and Y. S. Cho, "A new signal detection method for spatially multiplexed MIMO systems and its VLSI implementation," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 56, no. 5, pp. 399–403, May 2009.

[21] A. Maltsev, V. Pestretsov, R. Maslennikov, and A. Khoryaev, "Triangular systolic array with reduced latency for QR-decomposition of complex matrices," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2006, pp. 385–388.

[22] Z. Liu, K. Dickson, and J. V. McCanny, "Application-specific instruction set processor for SoC implementation of modern signal processing algorithms," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 52, no. 4, pp. 755–765, Apr. 2005.

[23] C. K. Singh, S. H. Prasad, and P. T. Balsara, "VLSI architecture for matrix inversion using modified Gram–Schmidt based QR decomposition," in *Proc. Int. Conf. VLSI Des.*, Jan. 2007, pp. 836–841.

[24] C. K. Singh, S. H. Prasad, and P. T. Balsara, "A fixed-point imple-mentation for QR decomposition," in *Proc. IEEE Dallas Workshop Des. Appl. Integr. Softw.*, Oct. 2006, pp. 75–78.

[25] S. Wang and E. E. Swartzlander, Jr., "The critically damped CORDIC algorithm for QR decomposition," in *Proc. IEEE Asilomar Conf. Signals Syst. Comput.*, Nov. 1996, pp. 908–911.

[26] H. Sakai, "Recursive least-squares algorithms of modified Gram–Schmidt type for parallel weight extraction," *IEEE Trans. Signal Process.*, vol. 42, no. 4, pp. 429–433, Feb. 1994.

[27] S.-F. Hsiao and J.-M. Delosme, "Householder CORDIC algorithms," *IEEE Trans. Comput.*, vol. 44, no. 8, pp. 990–1001, Aug. 1995.

Robert Chen-Hao Chang (S'91–M'95–SM'09) received the B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1987 and 1989, respectively, and the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, in 1995.

In 1996, he joined the faculty of the Department of Electrical Engineering, National Chung Hsing University, Taichung, Taiwan, where he is currently a Professor and where he was the Director of the Meng Yao Chip Center from 2000 to 2004, the Director of the Center for Research and Development of Engineering Technology, College of Engineering, from 2005 to 2006, and the Chairman of the Department of Electrical Engineering from 2006 to 2008. He has published more than 80 technical journal and conference papers. His research interests include power management IC design, low-power circuit design, and baseband circuit design.

Dr. Chang was a recipient of the Distinguished Teaching Award and the Outstanding Research Project Award from National Chung Hsing University in 2004 and 2006, respectively. He is a member of Tau Beta Pi. He has been an Associate Editor for the IEEE Transactions on VLSI Systems since 2010. He has been a member of the VLSI Systems and Applications Technical Committee and the Nanoelectronics and Gigascale Systems Technical Committee, IEEE Circuits and Systems Society, since 2004 and 2009, respectively. He also served as technical program committee member for many conferences.

Chih-Hung Lin received the B.S. and M.S. degrees in electrical engineering from National Chung Hsing University, Taichung, Taiwan, in 1996 and 2002, respectively, where he is currently working toward the Ph.D. degree in the ICs and system research group.

His research interest is VLSI architecture design for communication.

Chien-Lin Huang was born in Taiwan in 1985. He received the B.S. and M.S. degrees in electrical engineering from National Chung Hsing University, Taichung, Taiwan, in 2006 and 2009, respectively.

He is with the National Chip Implementation Center, Hsinchu, Taiwan.

Kuang-Hao Lin received the B.S. and M.S. degrees in electronics engineering from the Southern Taiwan University of Technology, Tainan, Taiwan, in 2001 and 2003, respectively, and the Ph.D. degree in electrical engineering from National Chung Hsing University, Taichung, Taiwan, in 2009.

After his graduate studies, he was with the SOC Technology Center, Industrial Technology Research Institute, Hsinchu, Taiwan. In 2009, he became an Assistant Professor with the Department of Electronic Engineering, National Chin-Yi University of Technology, Taichung. His research interests include multiple-input–multiple-output wireless communication systems, synchronization in digital communications, channel coding, and VLSI architecture design for communication.

Feng-Chi Chen was born in Taiwan in 1982. He received the B.S. degree from the National Changhua University of Education, Changhua, Taiwan, in 2005 and the M.S. degree in electrical engineering from National Chung Hsing University, Taichung, Taiwan, in 2007.

He is with the Information and Communications Research Laboratories, Industrial Technology Research Institute, Hsinchu, Taiwan.