Abstract—The singular value decomposition (SVD) and the QR decomposition (QRD) are two prominent matrix decomposition algorithms used in various signal processing applications. In the field of multiple-input multiple-output (MIMO) communication systems, the SVD and the QRD are employed for beamforming and for channel-matrix preprocessing for MIMO detection, respectively. In this paper, we describe a minimum-area matrix decomposition architecture that is programmable to perform QRD and SVD with variable precision and we investigate the associated design and implementation trade-offs. Our reference implementation achieves a hardware efficiency of up to 325 k SVDs/s/mm$^2$ and 1.92 M QRDs/s/mm$^2$ for complex-valued 4 × 4 matrices in 0.18 µm CMOS technology.

I. INTRODUCTION

Multiple-input multiple-output (MIMO) communication systems [1] constitute the basis for many upcoming wireless communication standards (e.g., IEEE 802.11n) and offer increased spectral efficiency (compared to single-antenna systems) by transmitting multiple data streams concurrently in the same frequency band. Matrix decomposition algorithms [2], such as the singular value decomposition (SVD) or the QR decomposition (QRD), have applications in various signal processing fields. The SVD, for example, is used in array processing or data compression, but can also be applied to MIMO systems in order to increase the system performance by the use of beamforming and power allocation. The QRD, for example, is a key prerequisite for many advanced MIMO detectors, such as the sphere decoder [3].

The SVD and the QRD mainly base on a specific sequence of Givens rotations [2]. CORDIC (coordinate rotation digital computer) algorithms have shown to be a suitable tool to efficiently perform Givens rotations in hardware [4]. Since the QRD only requires a subset of operations required for the SVD, an architecture which allows to compute the SVD would also provide all the functionality to compute the QRD. Hence, a programmable matrix decomposition architecture based on CORDIC arithmetic results in a single matrix decomposition unit (MDU) and is suitable for both decomposition algorithms.

Due to the relatively high computational complexity of the SVD, systolic arrays based on the Jacobi method have been proposed [4]–[6]. As illustrated in Fig. 1, systolic arrays lie on one end of the area/delay trade-off and are usually designed to achieve short computation time at the cost of large circuit area. However, in MIMO-OFDM systems [1] for example, multiple problems need to be solved concurrently, where the number of parallel tasks corresponds to the number of OFDM tones. The throughput of fast but large architectures (e.g., systolic arrays) is often difficult to match to an arbitrary number of problems, e.g., one systolic array might be insufficient in terms of throughput but two might exceed the available circuit area. Low-area architectures can be obtained by the use of time-sharing and lie on the other end of the area/delay trade-off (see Fig. 1). Ideally, time-shared architectures are equally efficient (in terms of area times the computation time) as systolic arrays and have the key advantage to be easily adaptable to individual throughput requirements by the use of parallel instantiation, i.e., the target throughput can be achieved by replication of a low-area instance. Additional area savings, while not reducing the decomposition throughput, can be obtained by iterative decomposition of each instance (see Fig. 1).

Contributions: We present reference VLSI implementation results of two MDUs optimized for MIMO systems. The low-area implementation mainly bases on CORDIC arithmetic and is able to perform the QRD and the SVD of complex-valued 4 × 4 matrices. In order to improve the overall efficiency, the involved design and implementation trade-offs are investigated in detail. We present two architecture variants: one unit has been optimized for throughput and the other offers enhanced flexibility by controlling a precision/throughput trade-off.

Outline: The remainder of this paper is organized as follows. In Sec. II the SVD and QRD algorithms are described and all required operations are identified. The time-shared matrix decomposition architecture is described in Sec. III and the associated design and implementation trade-offs are explored in Sec. IV. In Sec. V, we provide VLSI implementation results and conclude in Sec. VI.

![Fig. 1. Ideal impact of architectural transformations [7] on systolic arrays and time-shared architectures. The dashed box corresponds to the investigated design-space exploration for both MDUs by the use of iterative decomposition.](image-url)
II. MATRIX DECOMPOSITION ALGORITHMS

To identify all required arithmetic operations and the underlying computational sequences, the SVD and QRD algorithms are briefly summarized below. More details for both matrix decomposition algorithms can be found in [2].

A. Singular Value Decomposition

The SVD of a complex-valued $M \times N$-dimensional matrix $A$ is defined as [8]

$$A = U \Sigma V^H$$  \hspace{1cm} (1)

where $U$ and $V$ are complex-valued unitary matrices of dimension $M \times M$ and $N \times N$, respectively. The $M \times N$-matrix $\Sigma$ contains the real-valued singular values on its main diagonal, i.e., $\text{diag}(\Sigma) = \{\sigma_1, \sigma_2, \ldots, \sigma_r\}$, where $r = \min\{N, M\}$ and $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r$. The SVD procedure under consideration bases on the Golub-Kahan algorithm described in [2] and mainly performs the SVD in two phases:

1) Bidiagonalization: First, a memory is initialized with $M = \{I_M, A, I_N\}$ where $I_L$ stands for a $L \times L$ identity matrix. During the bidiagonalization phase, Givens rotations are successively applied to $A$ from the left-hand side (LHS) and from the right-hand side (RHS), such that the $M \times N$-dimensional inner matrix $A$ gets bidiagonal and real-valued (denoted by $B_0$) as illustrated in Fig. 2. All Givens rotations applied to $A$ from the LHS and RHS are applied to the corresponding identity matrices. The resulting unitary matrices are denoted by $U$ and $V^H$ and the memory content after the bidiagonalization phase corresponds to $M = \{U, B_0, V^H\}$ where $A = UB_0V^H$.

2) Diagonalization: The diagonalization phase consists of multiple diagonalization steps (indicated with $k$) and is illustrated in Fig. 2. Givens rotations are subsequently applied from the LHS and from the RHS to the bidiagonal matrix $B_k$ such that all off-diagonal entries $f_i$ (for $i = 1, 2, \ldots, r-1$) of $B_k$ become zero. The diagonalization phase is stopped whenever all $f_i$ are considered to be zero and all $d_1$ (for $i = 1, 2, \ldots, r$) correspond to the unordered singular values. In order to ensure convergence of the diagonalization phase and to reduce the overall computation time of the SVD, the first Givens rotation of each diagonalization step (indicated with RHS* in Fig. 2) is performed with a modified input vector $[x \ y]^T$, where $y = t_{12}$ and $x = t_{11} - \mu$ uses the Wilkinson shift [2]

$$\mu = a_n + c - \text{sign}(c)\sqrt{a_n^2 + b_{n-1}^2}$$  \hspace{1cm} (2)

with $c = \frac{1}{2}(a_{n-1} - a_n)$, $T = B_k^H B_k$, and the trailing non-zero sub-matrix of $T$ corresponds to

$$T(n-1 : n, n-1 : n) = \begin{pmatrix} a_{n-1} & b_{n-1} \\ b_{n-1}^* & a_n \end{pmatrix}.$$  \hspace{1cm} (3)

Analogous to the bidiagonalization phase, all Givens rotations are also applied to the corresponding unitary matrices such that finally, $M = \{U, \Sigma, V^H\}$ is the SVD in (1).

B. QR Decomposition

The QR decomposition of a $M \times N$-dimensional complex-valued matrix $A$ is defined as [8]

$$A = QR$$  \hspace{1cm} (4)

where $Q$ is a complex-valued $M \times N$-matrix with orthonormal columns and the upper-triangular $N \times M$-matrix $R$ has real-valued entries on its main diagonal. The QRD is performed in a similar fashion as the bidiagonalization phase of the SVD (cf. Fig. 2) where only the LHS Givens rotations are applied. This minor modification of the SVD algorithm results in an upper-triangular matrix $R$ with real-valued entries on its main diagonal. All Givens rotations applied from the LHS to $A$ are also applied to the unitary matrix $Q$ such that $M = \{Q, R, I_N\}$ corresponds to the QRD in (4).

III. VLSI ARCHITECTURE

In contrast to a systolic array implementation, we describe a low-area matrix decomposition architecture, which is designed to operate on complex-valued $4 \times 4$-dimensional matrices. The time-shared architecture is depicted in Fig. 3 and consists of tree main components described in the following.
A. Matrix Memory

The matrix memory provides storage for three complex-valued 4 × 4 matrices $M = \{M_1, M_2, M_3\}$, which is sufficient to store the result of an SVD and of a QRD (see Sec. II). A complex value in $M$ is stored at a single memory address, is 32 bits wide, and each real and imaginary part requires 16 bits. The matrix memory consists of a two-port 48 × 32 bit SRAM and requires 0.06 mm$^2$ in 0.18 μm CMOS technology. The matrix memory interface allows to read or write two different real or imaginary parts in at most two clock cycles.

B. Arithmetic Unit

In order to design a high-level VLSI architecture of the arithmetic unit (AU), Givens rotations, square-roots, multiplications, and additions/subtractions are required to compute the SVD and the QRD (cf. Sec. II). Givens rotations and the square root can efficiently be computed by CORDIC, whereas multiplications and additions/subtractions are computed in a multiply-accumulate (MAC) unit.

CORDIC Unit: CORDICs can efficiently compute two-dimensional rotations [9] by performing a series of micro-rotations with the aid of shifts and additions/subtractions (cf. Fig. 4). To keep the circuit area low, a single CORDIC is used by the means of time sharing and has been designed to support vectoring and rotation. A complex-valued Givens rotation is performed by three real-valued vectoring CORDICs (denoted by $C_1$, $C_2$, and $C_3$), i.e.,

$$
\begin{bmatrix}
C & C \\
R & C
\end{bmatrix}
\begin{bmatrix}
C_3 \\
R
\end{bmatrix}
\begin{bmatrix}
C_3 \\
R
\end{bmatrix}
\begin{bmatrix}
C_3 \\
R
\end{bmatrix}
\begin{bmatrix}
R \\
0
\end{bmatrix}.
$$

In order to perform the corresponding rotation on a complex-valued 2-dimensional vector, four rotation CORDICs are required: the first two ($C_1$ and $C_2$) rotate each complex entry independently, whereas the third and fourth CORDICs rotates the real and imaginary part of both complex values by $C_3$.

We emphasize that the square-root required in (2) can efficiently be computed with the CORDIC in vectoring mode, since $\sqrt{d^2 + b_{n-1}^2}$ corresponds to the nonzero result of the CORDIC with $[d b_{n-1}]^T$ applied to the input.

Multiply-Accumulate Unit: To compute the trailing sub-matrix of $T = B_R^T B_4$, as described in (3), a real-valued multiply-accumulate (MAC) unit has been instantiated. The multiplier can be switched off in order to perform additions or subtractions required in (2).

C. Instruction-Based Sequencer

The programmable MDU contains a micro-code controlled sequencer. This sequencer consists of a 64 × 20 bit instruction RAM (of size 0.04 mm$^2$ in 0.18 μm CMOS technology) that provides storage for 64 instructions. The finite state machine (FSM) decodes instructions, generates memory addresses, and provides control signals for the AU.

Instruction Set: The SVD and QRD mainly base on a specific rotation sequence applied to the matrices in $M$ (cf. Sec. III-A). To this end, a set of eight instructions has been defined. Four instructions are used to apply CORDICs from the left-hand side (LHS) or the right-hand side (RHS) to one complex-value or two real/imaginary-valued entries of $M_2$ and to update all other affected entries in $M$. One instruction is used to initialize the diagonalization phase of the SVD which subsequently performs all required diagonalization steps in a self-controlled manner. The remaining instructions are reserved to configure the number of CORDIC micro-rotations (see Sec. IV-A) or to control the program flow. The SVD of a complex-valued 4 × 4 matrix requires 27 instructions, whereas a QRD of equal size requires only 17 instructions.

SVD-Algorithm Modifications: To simplify the diagonalization phase of the SVD and to obtain a fixed throughput, the following modifications have been applied to [2]:

1) Off-diagonal entries of $B_k$ (see Fig. 2) are considered to be zero, whenever $f_i < 2^{-ε}$ for $i = 1, 2, \ldots, r - 1$.
2) Since the computational complexity of the diagonalization phase is data dependent, an early-stopping criterion has been introduced to obtain a fixed decomposition throughput. Whenever $k = K_{\text{max}}$, the diagonalization phase is stopped and the current $M$ is used as an estimate of (1).

Note that $ε$ and $K_{\text{max}}$ can be defined in the SVD-initialization instruction, which allows to reconfigure the arithmetic precision and the decomposition throughput at run time.

IV. IMPLEMENTATION TRADE-OFFS

To reduce the area of the MDU and to improve the overall efficiency, implementation trade-offs associated with arithmetic precision, circuit area, and throughput are investigated in the following.

A. Fixed-Point Implementation Trade-offs

The arithmetic precision of a fixed-point implementation is assessed by the bit error rate (BER) of an IEEE 802.11n-based MIMO-OFDM system with coded beamforming [10]. The baseband input-output relation of the wireless channel is $y = Hs + n$ where $H$ corresponds to the $M_R × M_T$-dimensional channel matrix, $s$ denotes the $M_T$-dimensional
transmit signal, \( \mathbf{n} \) the \( M_R \)-dimensional Gaussian noise vector, and \( \mathbf{y} \) the \( M_R \)-dimensional receive signal. Beamforming is simulated by computing the fixed-point SVD of the channel matrix \( \mathbf{H} \) and by transmitting \( \mathbf{s} = \mathbf{V} \mathbf{s} \). The receiver converts the input-output relation into \( M_T \)-single-input single-output channels

\[
\tilde{y}_i = \sigma_i \tilde{s}_i + \tilde{n}_i \quad \text{for} \quad i = 1, 2, \ldots, M_T
\]

where \( \tilde{y} = \mathbf{U}^H \mathbf{y} \), \( \sigma_i \) is the \( i \)th singular value, and the noise vector \( \tilde{\mathbf{n}} \) has equal statistics as \( \mathbf{n} \). Finally, a soft-output demapper produces reliability information for the subsequent soft-input channel decoder.

**Fixed-Point Implementation:** To convert the floating-point model in a fixed-point implementation, the floating-point SVD has been simulated to determine the threshold parameter \( \varepsilon \) and the required maximum number of diagonalization steps. Simulations have shown that setting \( \varepsilon = 7 \) and \( K_{\text{max}} = 7 \) does not result in a significant BER performance loss. Note that reducing \( \varepsilon \) or increasing \( K_{\text{max}} \) only increases the computational complexity of the SVD and does not improve the error rate performance. Further simulations have shown that a near-optimal BER is achieved by using 16 fractional bits within the CORDIC. Finally, the required number of micro-rotations in the CORDIC has been evaluated. Figure 5 shows the impact of micro-rotations to the BER of a beamformed MIMO-OFDM simulation\(^2\). At least 12 micro-rotations are required to achieve a BER less than \( 10^{-6} \). Reducing the number of micro-rotations results in a BER floor, which is not suitable for the scenario under consideration.

**Precision/Throughput Trade-off:** The number of micro-rotations in the CORDIC directly influence the arithmetic precision of the MDU (see Fig. 5) and has also a significant impact on the throughput. A lower number of micro-rotations

\(^2\)We consider a convolutionally encoded (rate 1/2, generator polynomials \([133, 171]_6\), constraint length 7, random interleaving) MIMO-OFDM system with beamforming [10], four transmit and receive antennas, 16-QAM (using Gray mapping), 64 tones, and soft-input Viterbi decoding. One codeblock corresponds to 1024 bits, a TGn type C [11] channel model is used, and perfect channel state information at the transmitter and receiver is assumed.

requires less clock cycles, which results in lowered computational complexity. We emphasize that the arithmetic precision requirements of a QRD for MIMO detection are usually lower than the precision for the SVD in the scenario considered above. As noted in [12], nine micro-rotations have found to be sufficient for a QRD and thus, it is beneficial to lower the precision in order to increase the decomposition speed. To this end, we propose a tunable arithmetic unit (denoted by AU II), where the number of micro-rotations is programmable, which allows to adjust the precision/throughput trade-off at run time. The maximum number of micro-rotations has been set to 12 to support sufficient precision for computing the SVD. An unroll factor of two has been chosen which allows to chose the number of micro-rotations from 12, 10, 8, and 6, depending on the application and precision requirements. Note that all BERs achievable by AU II are shown in Fig. 5.

**B. Area/Delay Trade-off**

Replication of a low-area unit can be used to achieve a given throughput. Lower area implies that the target throughput can be obtained more accurately. Hence, additional reduction in terms of area per decomposition unit, without a significant throughput decrease is highly desirable. To this end, iterative decomposition has been applied to the CORDIC unit to determine the optimum choice of the unroll factor. Since only the area of the AU is affected, the area/delay trade-off associated with the maximum number of micro-rotations and the CORDIC unroll factor is shown in Fig. 6. Two different MDUs have been designed: MDU I uses of the faster but less-flexible AU I (using 12 micro rotations and an unroll factor of six), whereas MDU II employs the slightly slower but configurable AU II (using an unroll factor of two with 12, 10, 8, and 6 micro-rotations), which allows to control the precision/throughput trade-off at run time. Note that the critical path is not only determined by the CORDIC, but also by the MAC unit. Thus, to align the critical paths of both units, one pipelining register has been inserted in the multiplier if the CORDIC unroll factor is less than four.
The fabricated ASIC in 0.18 μm technology is depicted in Fig. 7 and contains both MDUs and an unrelated design.

1) SVD: The VLSI implementation results for the SVD for each MDU are given in Tbl. I. Note that for highest precision (i.e., 12 micro-rotations), the first unit achieves a slightly higher throughput than the MDU II and only requires 0.04 mm² more area. However, the second unit is able to achieve 55% higher efficiency (in terms of SVDs per second per mm²) than MDU I by reducing the arithmetic precision down to six micro-rotations per CORDIC (see Fig. 5). At highest precision (i.e., using 12 micro rotations), MDU I and MDU II consume 160 mW and 106 mW, respectively.

2) QRD: The implementation results for a complex-valued QRD executed on each MDU are given in Tbl. II. At maximum precision (i.e., 12 micro-rotations which, however, is only desirable for the SVD computation), MDU I achieves the higher throughput than MDU II. Note that approximately six time more QRDs per second per mm² than SVDs are achievable. Reducing the precision of MDU II to six micro-rotations per CORDIC, allows to achieve 1.92 MQRDs/s/mm², which is 51% more efficient than the less-flexible MDU I. Hence, tuning the precision/throughput trade-off at run time can improve the overall efficiency of MDU II. Note that the power consumption increases by reducing the precision.

V. IMPLEMENTATION RESULTS

The fabricated ASIC in 0.18 μm technology is depicted in Fig. 7 and contains both MDUs and an unrelated design.

1) SVD: The VLSI implementation results for the SVD for each MDU are given in Tbl. I. Note that for highest precision (i.e., 12 micro-rotations), the first unit achieves a slightly higher throughput than the MDU II and only requires 0.04 mm² more area. However, the second unit is able to achieve 55% higher efficiency (in terms of SVDs per second per mm²) than MDU I by reducing the arithmetic precision down to six micro-rotations per CORDIC (see Fig. 5). At highest precision (i.e., using 12 micro rotations), MDU I and MDU II consume 160 mW and 106 mW, respectively.

2) QRD: The implementation results for a complex-valued QRD executed on each MDU are given in Tbl. II. At maximum precision (i.e., 12 micro-rotations which, however, is only desirable for the SVD computation), MDU I achieves the higher throughput than MDU II. Note that approximately six time more QRDs per second per mm² than SVDs are achievable. Reducing the precision of MDU II to six micro-rotations per CORDIC, allows to achieve 1.92 MQRDs/s/mm², which is 51% more efficient than the less-flexible MDU I. Hence, tuning the precision/throughput trade-off at run time can improve the overall efficiency of MDU II. Note that the power consumption increases by reducing the precision.

VI. CONCLUSION

We described design and implementation trade-offs of two programmable matrix decomposition units (MDUs), able to compute the SVD and the QRD. Low area is achieved through extensive use of time sharing of a single CORDIC unit. The low-area MDUs have been shown to be suitable for MIMO-OFDM systems, since they can be easily adapted to individual throughput requirements by the use of replication. One unit has been optimized for throughput, where the throughput of the second unit is tunable by reducing the arithmetic precision at run time. The programmability of both units allow to use the same architecture for different systems, which avoids the design of dedicated architectures for individual requirements.

REFERENCES


TABLE I

<table>
<thead>
<tr>
<th>MDU</th>
<th>I</th>
<th>II</th>
</tr>
</thead>
<tbody>
<tr>
<td>CORDIC micro-rotations</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Core area [mm²]</td>
<td>0.41</td>
<td>0.37</td>
</tr>
<tr>
<td>Maximum clock frequency [MHz]</td>
<td>133</td>
<td>272</td>
</tr>
<tr>
<td>Power Consumption [mW]</td>
<td>160</td>
<td>106</td>
</tr>
<tr>
<td>Maximum SVD time [ms]</td>
<td>11.57</td>
<td>13.29</td>
</tr>
<tr>
<td>Efficiency [SVDs/mm²]</td>
<td>240k</td>
<td>166k</td>
</tr>
<tr>
<td>Bit error rate floor [BER]</td>
<td>&lt; 10⁻⁶</td>
<td>&lt; 10⁻⁹</td>
</tr>
</tbody>
</table>

*Power consumption has been measured at the maximum clock frequency of the corresponding MDU with 1.8 V core supply.

TABLE II

IMPLEMENTATION RESULTS OF A COMPLEX-VALUED 4 × 4 QRD

<table>
<thead>
<tr>
<th>MDU</th>
<th>I</th>
<th>II</th>
</tr>
</thead>
<tbody>
<tr>
<td>CORDIC rots.</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>QRD time [μs]</td>
<td>1.92</td>
<td>2.82</td>
</tr>
<tr>
<td>QRDs/s/mm²</td>
<td>1.27 M</td>
<td>0.96 M</td>
</tr>
<tr>
<td>Power* [mW]</td>
<td>153</td>
<td>105</td>
</tr>
</tbody>
</table>

*Power consumption has been measured at the maximum clock frequency of the corresponding MDU with 1.8 V core supply.