PROCEEDINGS

Ninth Biennial
University/Government/Industry
Microelectronics Symposium

June 12-14, 1991
Melbourne, Florida
CAPE - VLSI Implementation of a Systolic Processor Array: Architecture, Design and Testing

Nariankadu D. Hemkumar
Kishore Kota
Joseph R. Cavallaro
Dept. of Electrical & Computer Engineering
Rice Univ., Houston, TX 77251
Phone (713).623.4000

Abstract

The SVD is an important matrix decomposition in many real-time signal processing, image processing and robotics applications. Special-purpose processor arrays can achieve significant speed-up over conventional architectures through the use of efficient parallel algorithms. The Cordic Array Processor Element (CAPE) is a single chip VLSI implementation of a processor element for the Brent-Luk Van Loan systolic array which computes the SVD of a real matrix. The array utilizes CORDIC (CO-ordinate Rotation Digital Computer) arithmetic to perform the vector rotations and inverse tangent calculations in hardware. A six-chip prototype of the processor has been implemented as TinyChips using the MOSIS fabrication service. Experience gained from designing the prototype helped in the design of integrated single chip version. The chip has been implemented on a 5600×6960µm die in a 2µm n-well scalable CMOS process.

1 Introduction

Many real-time signal processing applications require fast computation of matrix decompositions. Special-purpose systolic arrays [10] present an attractive architecture for implementing such factorization algorithms in hardware. The special-purpose nature of these arrays allows the optimization of system performance by exploiting the parallelism inherent in the problems. Systolic arrays, by definition, allow only near neighbor communication. The reduced path lengths and fixed fan-out allow design of large arrays with very little degradation in the clock speed.

The SVD is an important matrix factorization used in a variety of applications. The SVD exhibits better numerical stability due to the insensitivity to ill-conditioning or rank deficiency of matrices. However, the SVD is computationally intensive, requiring O(p^3), for a p×p matrix. For example, the SVD of an 8×8 matrix on a SUN 3/60 requires about 1 second of CPU time using the LINPACK routine DSVDC. As a representative real-time signal processing application, the inverse kinematics engine of a robot being constructed at Rice, requires the computation of the SVD of an 8×8 matrix in 400µs. With a square array of processors it is possible to reduce the time complexity for the SVD to O(p^log_2 p).

The SVD [8] of a p×p matrix M is

\[ M = U E V^T, \tag{1} \]

where U and V are p×p orthonormal matrices and Σ is a diagonal matrix of singular values. The columns of U and V are called the left and right singular vectors, respectively. The Jacobi method for computing the SVD is amenable to parallel and systolic computation. A sequence of orthonormal transformations are used to diagonalize a matrix M and the SVD is obtained by performing several sweeps of computation. A sweep consists of a sequence of 2-sided Jacobi rotations which annihilate every pair of off-diagonal elements at least once. It has been shown that O(log p) sweeps are required for convergence. If the sequence of rotations chosen is the parallel ordering [1], then it is possible to implement a sweep consisting of p( p−1)/2 2-sided rotations in O(p) time on a square array of p×p processors.

The Brent, Luk and Van Loan systolic array [1] uses a square array of p×p processors for SVD of a p×p matrix, where each processor has the ability to compute the SVD of a 2×2 matrix. A 2×2 SVD can be described as

\[ R(θ)^T \begin{bmatrix} a & b \\ c & d \end{bmatrix} R(θ) = \begin{bmatrix} ψ_1 & 0 \\ 0 & ψ_2 \end{bmatrix}, \tag{2} \]

where θ₁ and θ₂ are the left and right rotation angles, respectively. The rotation matrix is

\[ R(θ) = \begin{bmatrix} \cos θ & -\sin θ \\ \sin θ & \cos θ \end{bmatrix}, \tag{3} \]

and the input matrix is

\[ M = \begin{bmatrix} a & b \\ c & d \end{bmatrix}. \]

The angles θ₁ and θ₂ are given by the relations

\[ \begin{align*}
θ₁ + θ₂ &= \tan^{-1} \frac{c + b}{d - a}, \\
θ₁ - θ₂ &= \tan^{-1} \frac{c - b}{d + a}.
\end{align*} \tag{4} \]

For a p×p matrix, the Jacobi rotation is given by

\[ J = \begin{bmatrix} R(θ) & 0 \\ 0 & I \end{bmatrix}, \tag{5} \]

where R(θ) is a 2×2 rotation and I is a (p−2)×(p−2) identity matrix. A 2-sided rotation for a p×p matrix consists of a left and a right Jacobi rotation. Any pair of off-diagonal elements can be annihilated through a 2-sided rotation where the rows and columns of the Js are appropriately permuted. In a square systolic array (Figure 1), the transformations to annihilate the off-diagonal elements of the p×p diagonal processors are independent of the sequence in which the elements are annihilated, and hence can be evaluated concurrently. The angles corresponding to these transformations are computed by the diagonal processors using (4). The rest of the matrix is then updated by systolically propagating the left angles horizontally and right angles vertically to transform all the 2×2 submatrices residing in the various processors. The data is then permuted to present a new set of off-diagonal elements at the diagonal. In the array, application of the previous set of transformations and computation of a new set of rotations is pipelined. This leads to waves of activity originating at
the main diagonal and systolically propagating towards the northeast and southwest corners of the array. The data exchange algorithm, called the parallel ordering, has the property that the next set of data items to be annihilated is always available in the diagonal neighbors of the main diagonal. Thus all data exchanges are near-neighbor.

2 CORDIC Arithmetic

In special-purpose VLSI processors, an efficient design with optimal area and time complexity can be obtained by using some arithmetic techniques which map the desired computations to hardware. The principal computations involved in the SVD algorithm are vector rotations and inverse tangent calculations which can be implemented efficiently using the CORDIC technique. Other operations required are addition, subtraction and divide by two which can utilize the same functional units as CORDIC.

The Coordinate Rotation Digital Computer (CORDIC) technique was initially developed by Volder [14] as an algorithm to solve the trigonometric relationships that arise in navigation problems. Involving only a fixed sequence of additions or subtractions, and binary shifts (Figure 2), this scheme was used to quickly and systematically approximate the value of a trigonometric function or its inverse. This algorithm was later unified for elementary functions by Walther [15].

The basic iteration of the circular mode of operation in CORDIC implements a $2 \times 2$ Jacobi rotation (3) of a vector and can be expressed as

$$
\begin{align*}
\bar{x}_{n+1} &= x_n + \delta_n y_n 2^{-n} \\
\bar{y}_{n+1} &= y_n - \delta_n x_n 2^{-n} \\
\end{align*}
$$

where,

$$
\delta_n = \tan^{-1}(2^{-n})
$$

The parameter $\delta_n$ is chosen at each iteration to either force $z_0$ to 0 (Rotation or Z reduction) or $y_0$ to 0 (Vectoring or Y-reduction). Each iteration corresponds to a vector rotation of $[x_n, y_n]^T$ through an angle $\alpha_n$ with the direction of rotation determined by $\delta_n$, followed by a scaling by a factor $1/\cos \alpha_n$. The variable $x_n$ is either used to store the initial angle or accumulate the resulting angle. Z-reduction corresponds to an anti-clockwise rotation of the vector $[x_0, y_0]^T$ through an angle $z_0$. A post-processing step is required to eliminate the constant scale factor accumulated at each iteration. Y-reduction is used to compute $\tan^{-1}(y_0/x_0)$.

The CORDIC technique needs modification in order to be used in the SVD processor. In y-reduction, traditional CORDIC utilizes the sign of $y_n$ at each iteration to determine the $\delta_n$ for the next iteration. In the modified CORDIC algorithm [16], the exponents of the sign bits of $x_n$ and $y_n$ is used to determine this sign. Otherwise, in order to allow CORDIC to converge for vectors with a negative z component, a pre-processing step is necessary to invert the signs of both $z_0$ and $y_0$ if $z_0$ is negative.

In the computation of $\tan^{-1}(y_0/z_0)$, the actual magnitude of $z_0$ and $y_0$ should not affect the accuracy of the computed angle. However, in traditional CORDIC small values for $z_0$ and $y_0$ result in a loss of precision greater than that for larger magnitudes. Thus normalization of the values for inverse tangent calculations can produce better results. A straightforward way to implement this would be to use a pre-processing step in which $z_0$ and $y_0$ are shifted left by the maximum amount which will not cause an overflow in subsequent CORDIC iterations. A single clock cycle implementation of this scheme would require a barrel shifter with $O(n^2)$ area complexity and a leading-zero encoder with $O(2^n)$ complexity. In this paper a novel scheme [9] for fixed point CORDIC which allows the normalization to be made part of the CORDIC iterations without any time penalty, is presented. The scheme modifies the CORDIC iterations with a small left shift of the variables $z_n$ and $y_n$ if their magnitude is too small. Within a few iterations the variables are normalized and do not need any more shifts. This scheme achieves results identical to the scheme which uses a pre-processing step. The above technique reduces the hardware complexity for normalization from $O(2^n)$ to $O(n \log n)$ where $n$ is the bit precision. This reduction was achieved by using a shifter which implements less than $\log n$ shifts and a leading-zero encoder for $\log n$ bits.

Since the basic computations required for a $2 \times 2$ SVD map directly to the primitives provided by CORDIC, a $2 \times 2$ SVD processor reduces to the following algorithm [6]:

Algorithm CORDIC SVD( )

begin

Use CORDIC Y-Reduction to find rotation angles;

Use CORDIC Z-Reduction to transform the $2 \times 2$ matrix;
end

Figure 1: A linear array of DSP chips interfaced to a square array of special-purpose CAPE chips for an 8-degree of freedom robot.

Figure 2: Block Diagram of CORDIC.
3 Design of the SVD array

Figure 1 shows the inter-connection of processors to implement the systolic array. The square array of special-purpose SVD processors is connected to a linear array of general purpose DSP chips which are used for general purpose computations. The array serves as a co-processor for matrix factorizations, in particular, the SVD. The CAPE chip has been designed to allow the array to be used for matrix factorizations other than the SVD. In particular, it would be possible to compute the QR decomposition and Eigenvector decomposition of a symmetric matrix on the array. These are some of the computations which are encountered in many engineering problems.

The CORDIC systolic processor array (CSPA) is a mesh-connected array of processors with additional diagonal links to facilitate data exchanges. Horizontal and vertical links are used to systolically propagate the left and right rotation angles, respectively.

CAPE is a single chip implementation of the CORDIC SVD processor [6]. The SVD algorithm discussed in § 1 and the CORDIC algorithm described in § 2 form the basis for the design of the CORDIC-SVD processor. The processor was designed in two phases. Initially, a six chip prototype of the processor was designed. Each component chip was fabricated as a “TinyChip” from the MOSIS fabrication facility. Implementation of the processor as a set of chips provided controllability and observability which was essential at that stage of design. The prototype served as a means of exhaustively testing every aspect of the design which is not possible with simulation. Digital logic external to the chips could be used to correct inconsistencies in the interaction between the sub-units. The overall design was then verified in spite of minor design errors.

The second phase involved designing a single chip version of the processor. This chip utilized many of the basic blocks from the six-chip prototype in an enhanced architecture. The single chip version utilized better layout techniques using higher level design tools. Some of the lower level VLSI layout issues are discussed in § 6.

4 Architecture of the Prototype

Figure 2 shows an implementation of a 16-bit fixed point CORDIC processor. The xy-chip implements the x and y data paths of a single CORDIC module. A previous implementation of the xy-chip was reported by Cavallaro [3]. Each path consists of a 16-bit register, a 16-bit barrel shifter implementing all right shifts up to 15, a 16-bit ripple-carry adder with associated temporary latches at the inputs, and a data switch to implement the cross-coupling implied by the x and y CORDIC iterations. Both CORDIC iterations and scale factor correction iterations are possible on the data stored in the registers. The absence of any on-chip control allows an external controller complete flexibility of the number of iterations as well as the nature of each iteration. Read and write access to the x and y registers is possible independent of the computation. Figure 4 shows a die photograph of the xy-chip.

4.1 Design of the XY-Chip

The z-chip complements the xy-chip to form a complete 16-bit fixed point implementation of the CORDIC module. It implements the z data path which consists of a 16-bit z-register, a carry-lookahead adder and a ROM table of 16 indices. It also implements a controller to provide the control signals for the z data paths in the xy-chip and the z datapath in the z-chip. The controller allows use of the CORDIC module in five modes: z-reduction, y-reduction, scale factor
correction, single add and subtract, and divide-by-two. These are the only operations required in the computation of the SVD. Figure 5 shows a die photograph of the z-chip.

4.3 Design of the Intra-chip

The intra-chip controls data movement between a register bank and the two CORDIC modules and sequences the primitives provided by the z-chip to compute the SVD of the $2 \times 2$ matrix stored on the processor. It allows operation in three modes corresponding to processor behavior in three different positions in the array. If the processor is on the main diagonal of the array, it computes the left and right angles and then uses these angles to diagonalize a $2 \times 2$ submatrix. Any off-diagonal processor only transforms a $2 \times 2$ submatrix with the angles received from the previous processors. However, if a processor is diagonally a multiple of three away from the main diagonal, in addition to the transformation, a delay equal to the time it takes to compute the angles is included. This delay is necessary to maintain synchronization between all the processors due to the wavefront nature of the array. These three modes allow the same processor to be used in any position of the array in spite of the required heterogeneity. The aforementioned storage is a register bank consisting of ten registers implemented in the intra-chip. A 16-bit bus interconnects these registers and the two CORDIC modules. Figure 6 shows a die photograph of the intra-chip.

4.4 Design of the Inter-chip

The inter-chip contains the controller which exercises the highest level of control in the processor. As a communication agent, it implements the systolic array control for the SVD processor. The parallel ordering data exchange [1] which provides a new pair of data elements at the main diagonal to be annihilated, is implemented in this chip. A set of six shift registers is used as communication buffers to store a copy of the data to be exchanged with the other processors while the inter-chip operates concurrently on its own private copy of the data.

Interprocessor communication is through a set of eight serial links. Four of these links are hardwired to exchange data with the diagonal neighbors. Four others links serve to systolically propagate the left angles horizontally and and right angles vertically. Since the SVD algorithm is compute bound, the bandwidth provided by the serial links is sufficient for the data interchange. In addition, pin limitations prohibit implementation of parallel ports to communicate with neighboring processors. Data is exchanged synchronously between processors. Data is shifted out to the neighboring processor and simultaneously shifted in from it. Synchronization in the array is of utmost importance since no other form of synchronization is provided. This is in accordance with the definition of systolic arrays [10] which emphasizes exploiting the special-purpose nature of the array to eliminate the overhead of synchronization. In the prototype, pin limitations forced a serial link for communication of data with the intra-chip. The inter-chip controller also allows initialization of the array by loading the matrix data. Figure 7 shows a die photograph of the inter-chip.
5 Architecture of CAPE

The CORDIC array processor element (CAPE) is a single chip implementation of the CORDIC SVD processor. This chip incorporates elegant solutions to the problems encountered in the six-chip prototype. CAPE also includes additional details necessary to construct a working systolic array. CAPE essentially consists of the same basic cells as the prototype. Many of the architectural enhancements were made to reduce the communication required within the chip. A completely new loading scheme was added to facilitate easy interface with a variety of hosts. The next few paragraphs describe the details of the improvements.

A study of the prototype CORDIC units indicated that better numerical accuracy could be achieved with some modifications to the modules as described in §2. To improve the numerical accuracy of the inverse tangent calculations, a normalization shifter was included in each of the z and y data paths. These shifters are activated during y-reduction by two leading-zero encoders which monitor the magnitude of the z and y data. Simulations indicate that this scheme provides several extra bits of accuracy in many cases.

The CORDIC modules in the prototype converge only for vectors in the first and fourth quadrants while operating in the y-reduction mode. The modified scheme uses the exclusive-or of the sign bits of the z and y registers to control the next iteration. This effectively increases the range of convergence for the CORDIC modules. In addition, the above scheme adds symmetry to the z and y data paths since they cannot be distinguished externally. This allows the same CORDIC module to rotate a vector in either clockwise or counterclockwise direction. Some data movement is rendered unnecessary by the above method, for example, if the data to be used in the next computation already resides in the z and y registers in the wrong order, since the definition of the z and y registers can be interchanged, no data movement is required.

The z-chip was modified to allow mapping the z, y, and z registers as part of the register bank. This allows reuse of the data left in the registers by a previous computation which is important when attempting to cut down the communication overhead.

Figure 8: Internal Architecture of CAPE

A study of the prototype SVD processor showed that there was a large overhead incurred in moving data to the appropriate registers between computations. A double bus architecture, as shown in Figure 6, was used to minimize this communication. A separate CORDIC module is hardwired to communicate on each bus. The set of six registers is split into two banks which can communicate with either bus allowing simultaneous access to two different registers. The number of registers required in the architecture was reduced from ten to six due to mapping of the CORDIC registers as general purpose registers. The double bus allows concurrent data movement to the two CORDIC modules which reduces communication cycles by half. Using the double bus architecture did not cause an increase in the area of the chip. This was a result of the reduction in the size of the intra-chip controller to half its original size. This corresponds to a reduction in the number states due to greater concurrency provided by the architecture.

Special cases that occur in the SVD computation require extra hardware. Two zero comparators were included to detect the case \( \tan^{-1}(0/0) \). This is an invalid computation on the CORDIC modules which return a code equivalent of the floating-point NaN. The SVD algorithm allows such computed angles to be replaced by zero. The two zero comparators compare the data on the bus with zero and return a control signal to the intra-chip controller to handle this special case. A literal constant zero can be driven onto either bus and hence into any register. This allows forcing a register to store a zero to handle special cases.

The inter-chip has been modified to handle an improved array loading scheme. The prototype did not make a special provision to efficiently load the array. Since CAPE is to be used in a systolic array in a real system, the interface with the host has been refined. A parallel port has been included along with the requisite handshake signals. The controller implements an asynchronous loading scheme which allows interfacing with a host running at an independent clock rate.

The prototype inter-chip was pin limited and hence forced to communicate with the intra-chip through a serial link. This restriction is no longer valid in the single chip implementation. Consequently, this serial link was replaced with an internal parallel bus which reduces the time to exchange data items between the intra-chip and the inter-chip. The rest of the inter-chip was redesigned with only minor changes.

6 VLSI Implementation

The six-chip prototype was implemented by dividing the system into four logical sub-units. This approach inherently allows division of labor. The different design groups involved could work in parallel with the minimum of interaction once the component sub-units had been identified and their interactions defined. Different versions of the sub-units were implemented to explore design alternatives. Combining the verified sub-units to obtain a working prototype was then a simple task. A structured VLSI design approach was used.
following the Mead and Conway methodology [11]. Static complementary (CMOS) logic was used to simplify design. In addition, a formal two-phase clocking strategy was imposed [7, 8]. The individual units were developed using the Magic [12] layout editor and other VLSI tools from the University of California, Berkeley. Layout was done in a bit-slice manner. Programmable Logic Arrays (PLAs) were generated automatically from a high level description. Thus, a semi-custom design style was adopted. Switch-level simulation was performed using eain and irsim. Longest path delays were identified with crystal. Detailed simulation at the circuit level was performed for critical cells using spice. Chips were designed using the scalable CMOS technology from the MOSIS fabrication service. The single chip utilized the better routing facilities available with OCTTOOLS, also from Berkeley [10]. The design times for future versions of CAPE can be considerably sped up by using a symbolic layout style using these tools. The TinyChips were fabricated on 2200 × 2200μ die using a 2µ n-well process. The transistor count for the different chips varied between 4500 and 9000 transistors. The single chip version CAPE was designed on a 6000 × 5600μ die and contained about 26000 transistors. A plot of this chip is shown in Figure 9.

PC-based high speed test equipment, Omnilab from Orion Instruments [6], was used in verifying operation of the chips. This tester allows testing up to a frequency of 17 MHz which corresponds to a clock frequency of 4.25 MHz for the chip, due to the two-phase clocking strategy. All the components operate at the maximum test frequency. Exhaustive functional testing of some of the components was done using the parallel ports provided on an interface card on an IBM PC-AT. A C-language library of routines was created to use the CORDIC module as a co-processor for the IBM PC. This allowed an exhaustive test of its numerical stability.

7 Current Work and Summary

The CAPE project, although research oriented, provided sufficiently diverse and challenging VLSI projects that could be implemented over the period of an academic semester and within the confines of a course in VLSI and CAD. The six chip prototypes have been built and tested exhaustively. A single chip version of the processor has been sent for fabrication. The principal funding for fabrication of the sub-units came from the VLSI teaching grants provided by NSF. The integration of the sub-units into a single chip was funded through a research initiation grant from NSF.

Future versions of CAPE can implement floating point CORDIC units. Although only the SVD of a real matrix has been implemented, an architecture for the SVD of a matrix with complex data has also been developed. This array can be developed starting from the current design. An architecture has been developed to compute subsidiary matrices like U and V matrices in (1). Additional goals include implementation of a fault-tolerant array. These versions can benefit from a shorter design cycle due to better place and route VLSI tools like OCTTOOLS.

Acknowledgments

This work was supported in part by the National Science Foundation under Research Initiation Award MIP-8909498.

References


