VLSI DESIGN
2000, Vol. 11, No. 4, pp. 353–361
Reprints available directly from the publisher
Photocopying permitted by license only

© 2000 OPA (Overseas Publishers Association) N.V.
Published by license under
the Gordon and Breach Science
Publishers imprint.
Printed in Malaysia.

Design and Analysis of Digital Ratioed Compressors for Inner Product Processing*

CHUA-CHIN WANG[†], CHENN-JUNG HUANG and PO-MING LEE

Department of Electrical Engineering, National Sun Yat-Sen University, Kaohsiung, Taiwan 80424

(Received 5 June 1999; In final form 10 September 1999)

Inner product calculations are often required in digital neural computing. The critical path of the inner product of two binary vectors is the carry propagation delay generated from individual product terms. In this work, two architectures to arrange digital ratioed compressors are presented to reduce the carry propagation delay in the critical path. Besides, the carry propagation delay estimation of these compressor building blocks is derived and compared. The theoretical analysis and Verilog simulation both indicate that one of the compressor building blocks we present here might offer a sub-optimal solution for the basic building blocks used in digital hardware realization of the inner product computation.

Keywords: Ratioed compressors; Digital neural computing; Inner product computation

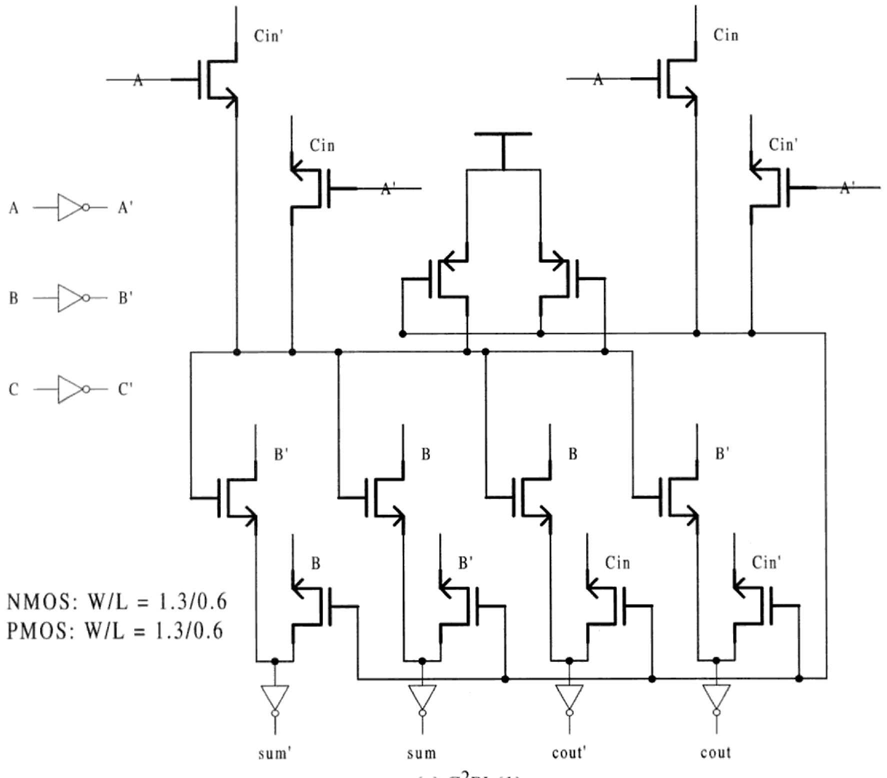
1. INTRODUCTION

Many efforts have been thrown on the realization of neural networks mainly owing to their attractive pattern recognition features, [1, 2]. In the computation of neural networks, the inner product of two vectors might be one of the most frequently used mathematical operations. Unavoidably the carry propagation will occur if the neural networks are dedicated for either discrete or digital signals. For instance, the recall of pattern pairs stored in discrete bidirectional associative memory (BAM) needs to compute a summation in the form as $Y = \mathrm{th}\left(\sum_{i=1}^{n} Y_i \cdot (X_i \cdot X)\right)$ where X is the input pattern, Y is the output pattern, X_i's and Y_i's are stored pattern pairs, and th( ) is a threshold function. Notably, the components of every vector are either bipolar or binary. If n is large in the above calculation, then the carry propagation of the inner product of the vectors will likely become the critical delay of the entire neural computing.

*This research was partially supported by National Science Council under grant NSC 88-2219-E-110-001. Based on "A comparison of two alternative architectures of digital ratioed compressor design for inner product processing", by C.-C. Wang, C.-J. Huang, and P.-M. Lee, which appeared in *Proceedings of 1999 Inter. Symp. on Circuits & Systems*, Vol. I, 161–164, Orlando USA, June, 1999. ©1999 IEEE.

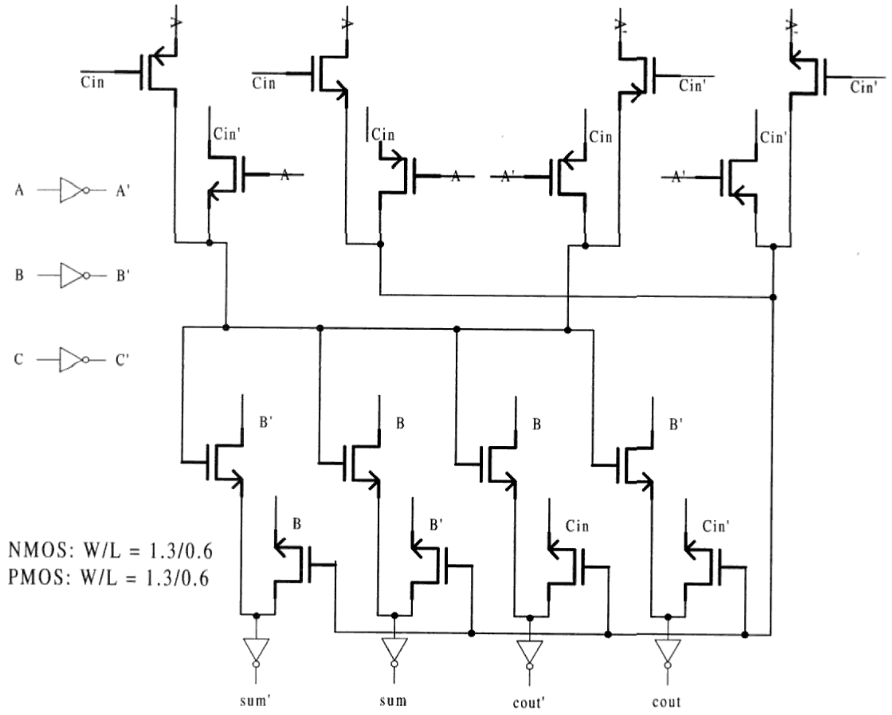
[†] Corresponding author. Tel.: 886-7-525-2000, ext. 4144, Fax: 886-7-5254199, e-mail: ccwang@ee.nsysu.edu.tw

Since neural computing is composed of mass amount of inner product calculations, the demand of shortening the delay therewith becomes urgent. Many high-speed logic design styles have been announced. However, these logics suffer from different difficulties. For example, domino logic [3] can not be non-inverting; NORA [4] has the charge sharing problem; all-N-logic [5] and robust single phase clocking [6] cannot operate correctly under clocks with short rise time or fall time, which can not be easily integrated with other part of logic design; single-phase logic [7] and Zipper CMOS [8] contain slow P-logic blocks. Though Zhang *et al.* [9] proposed a design of compressor to fix such a problem by employing a so-call C^2PL (complex CPL), several physical design factors are not fully considered or implemented. First, the sizes of the NMOS transistors for pass logics are impossible to be minimal. Second, the driving inverters' sizes have to be properly tuned. Third, the original design of [9] not only gives a poor fan-in and fan-out capability, but also produces very asymmetrical rise delay and fall delay which will very much likely cause glitch hazards and unwanted power consumption. Fourth, no further analysis on reduction of carry propagation delay is performed. In this paper, two alternative architectures of the digital ratioed compressors building



(a) C^2PL(1)

FIGURE 1 Schematic diagram for C^2PL 3-2 compressor in original design.



(b) $C^2PL(2)$

FIGURE 1 (Continued).

blocks based on the 3-2 compressors are presented, where the problems mentioned above are all resolved. An analytical form of carry propagation delay estimation for these two architectures is also derived. At last, the HSPICE and Verilog simulation results are also presented to verify the correctness of our observation.

2. FRAMEWORK OF RATIOED COMPRESSOR BUILDING BLOCKS

2.1. Basic Compressor Building Block Design

A 3-2 compressor is basically a full adder. The equations of a full adder are shown as follows:

$$\begin{aligned} S &= (a \oplus c)b' + (a \oplus c)'b = Fb' + F'b \\ C &= (a \oplus c)b + (a \oplus c)'c = Fb + F'c, \end{aligned} \tag{1}$$

where F denotes $(a \oplus c)$. The feature of such a compressor is that the output represents the number of 1's given in inputs.

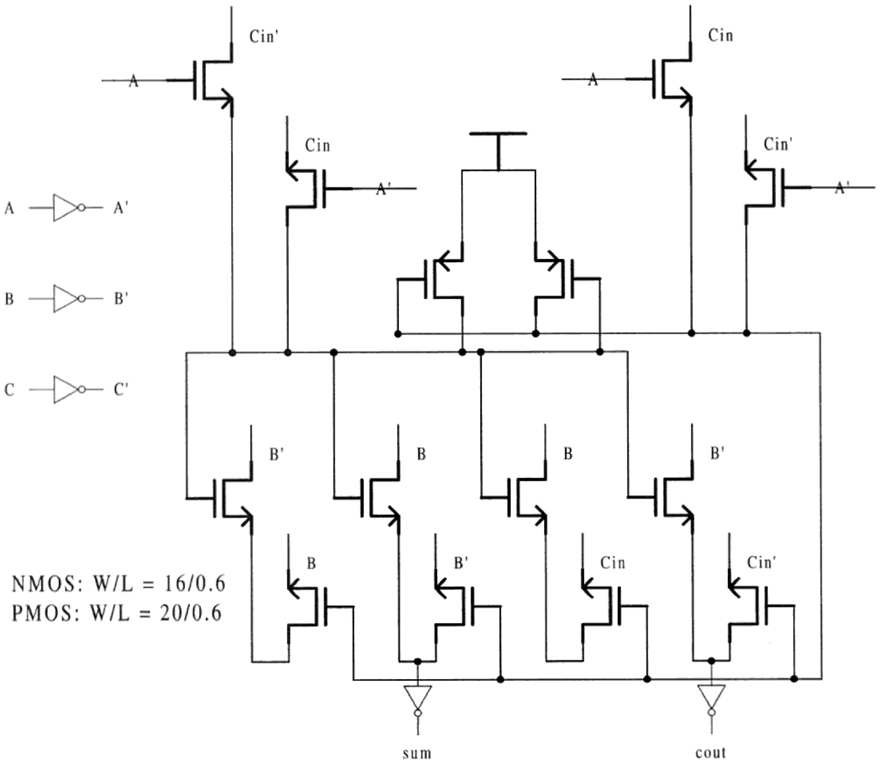
2.2. Ratioed 3-2 Compressor Design

Though a 3-2 compressor can be realized by a full adder, and Zhang *et al.* [9] proposed a C^2PL design for 3-2 and 7-3 compressors, several design issues as addressed in Section 1 are still ignored in their work. Figure 1 shows the schematic diagrams for the two types of 3-2 compressors based on complex complementary pass-transistor logic (C^2PL) proposed in [9]. We use TSMC 0.6 μm 1P3M technology to re-design the 3-2 compressors, and the schematic diagrams for the ratioed 3-2 compressors are shown in Figure 2. In Section 3 of this paper, we will demonstrate that the re-designed 3-2 compressor circuits will overcome all of the problems mentioned in Section 1.

2.3. A Primitive Architecture of Digital Ratioed Compressors

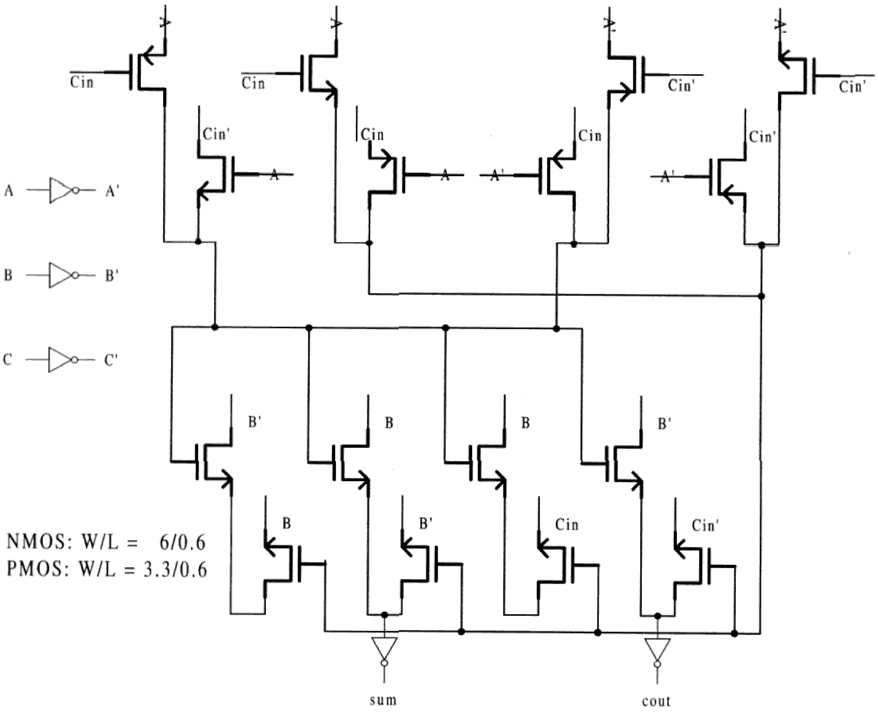
A 7-3 compressor building block can be constructed by cascading four 3-2 compressors as shown in Figure 3. A 15-4 compressor building block can also be formed similarly with two 7-3 compressors and two 3-2 compressors, as shown in Figure 4. Based on this design methodology, a $(2^n - 1)$-to-n compressor can be composed of two $(2^{n-1} - 1)$-to-$(n-1)$ compressors and $(n-1)$ 3-2 compressors.

Since the total delay of such design is approximately proportional to the count of 3-2 compressors that the critical path resides, we assume D_n denotes the count of 3-2 compressors when $2^n - 1$ bits are fed into the $(2^n - 1)$-to-n compressor block. By observing the structure of the compressor blocks, we can deduce D_2, D_3, and D_n as



(a) C^2PL(2)

FIGURE 2 Schematic diagram for re-designed C^2PL 3-2 compressor.



(b) C^{2}PL(2)

FIGURE 2 (Continued).

follows:

$$\begin{aligned} D_2 &= 1 \\ D_3 &= 1 = 1 + 2 = 2 + D_2 \\ D_n &= n - 1 + D_{n-1}, \quad n \geq 3. \end{aligned} \tag{2}$$

By solving the above recurrence relation, we obtain

$$D_n = \frac{n(n-1)}{2}. \tag{3}$$

Apart from the delay for the single building block, we have to count in the processing stages needed for summing individual inner product terms. The numbers of processing stages is roughly estimated as $\ln(n/M)/\ln(n/(2^n-1))$, where n denotes the total bits of the basic building block output, and M represents the bit count of data inputs.

Therefore, the count of 3-2 compressors when M bits are fed into the (2^n-1)-to-n compressor building blocks can be shown as follows:

$$D_{M,n} = \frac{\ln(n/M)}{\ln(n/(2^n-1))} \cdot \frac{n(n-1)}{2}. \tag{4}$$

2.4. The Systolic-like Architecture of Ratioed Compressor Building Blocks

The second architecture presented in this work to improve the carry propagation delay of the critical paths is shown in Figure 5. This architecture, inspired by the design methodology of systolic arrays, consists of parallelized 3-2 compressor building blocks only at every processing stage.

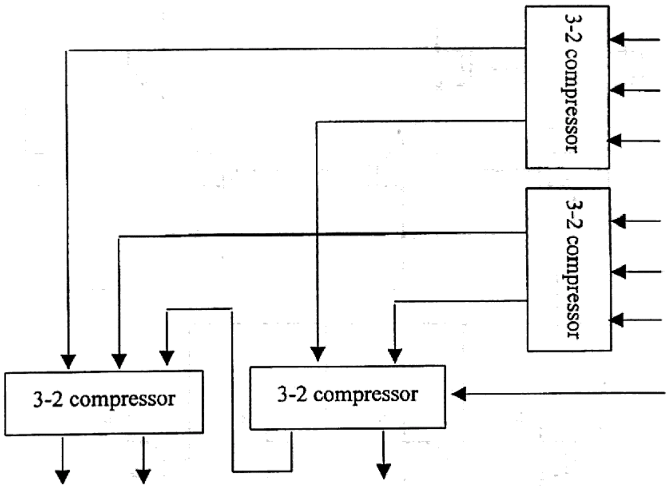


FIGURE 3 A 7-3 compressor building block.

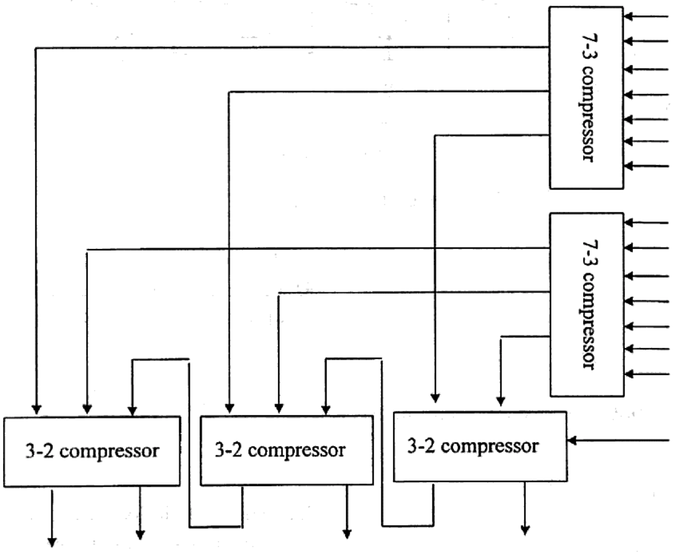


FIGURE 4 A primitive architecture of 15-4 compressor building block.

Although it is difficult to derive the analytical form of total delay of (2^n-1)-to-n compressors for systolic-like architecture, the upper bound for the delay of (2^n-1)-to-n compressors can be still computed in light of the result given in Eq. (4); *i.e.*,

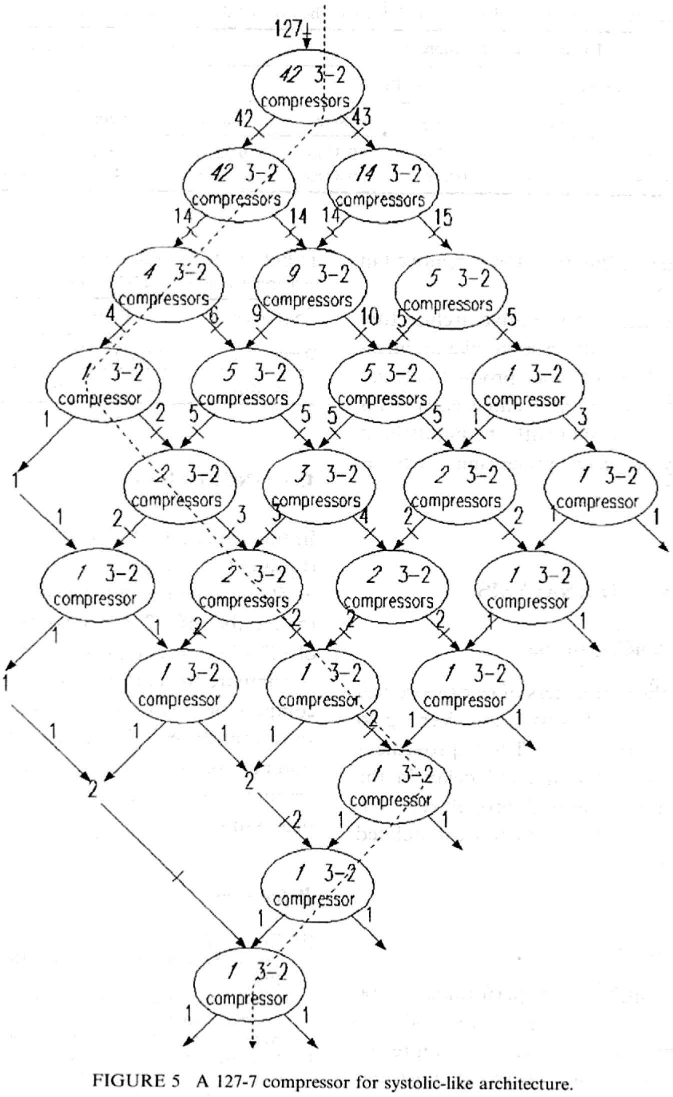


FIGURE 5 A 127-7 compressor for systolic-like architecture.

$$D_n = \left\lceil \frac{\log((2^n-1)/2)}{\log(3/2)} \right\rceil + c < \left\lceil \frac{(n-1)\cdot \log 2}{\log(3/2)} \right\rceil + c < (n-1)\cdot \left\lceil \frac{\log 2}{\log(3/2)} \right\rceil < 2\cdot(n-1), \tag{5}$$

where c is a small integer which is used to offset the bias between the estimated and the correct value of the total delay introduced in Eq. (4). Note that c is much smaller than the first dominant term embraced with ceiling function in Eq. (5), thus it

TABLE I The comparison of rise delay and fall delay in the original design and the re-designed 3-2 compressor

Circuits	The original 3-2 compressor				The re-designed 3-2 compressor			
	C^2PL (1)		C^2PL (2)		C^2PL (1)		C^2PL (2)	
Delay	Carry	Sum	Carry	Sum	Carry	Sum	Carry	Sum
Rise delay	0.26 ns	0.31 ns	0.42 ns	0.35 ns	0.32 ns	0.36 ns	0.41 ns	0.34 ns
Fall delay	0.87 ns	0.83 ns	0.87 ns	0.87 ns	0.24 ns	0.43 ns	0.39 ns	0.42 ns

can be ignored when ceiling function is removed in the above equation.

Comparing with the first primitive architecture presented in Section 2.3, the systolic-like architecture improves the delay of inner product calculation from $O(n^2)$ to $O(n)$. Apparently this outperformance is associated with the parallelized computing ability at each processing stage as shown in Figure 5.

3. SIMULATION AND ANALYSIS

3.1. Re-designed Building Blocks

In order to verify the correctness of our theoretical analysis, we tend to use HSPICE and Verilog to conduct a series of simulations. The improvement of asymmetrical rise delay and fall delay in the original design can be illustrated through HSPICE simulations. The simulation results are tabulated as shown in Table I.

3.2. Delay Simulations

The Verilog simulations are performed 20000 iterations for the first architecture and the systolic-like architecture of 127-7 compressor building blocks, respectively. Table II illustrates the comparison of carry propagation delay for the two architectures of 127-7 compressor building blocks when they are fed with 127 data inputs summation.

The results demonstrate that the systolic-like architecture of digital ratioed compressors indeed lead the least carry propagation delay.

TABLE II The comparison of carry propagation delay for the two architectures of 127-7 compressors

Circuits	Form I		Form II	
Delay	C^2PL (1)	C^2PL (2)	C^2PL (1)	C^2PL (2)
Delay (ns)	10	9	4	5

4. CONCLUSION

In this paper a re-designed ratioed 3-2 compressor is presented to correct several problems appearing in Zhang's work in [9]. The equations for counting the number of 3-2 compressors in the critical path of (2^n-1)-to-n compressors are derived and used to compare the performance of two digital ratioed compressor architectures. Our simulation results show that the systolic-like architecture gives a sub-optimal performance through the parallelized arrangement of 3-2 compressors at each stage of processing.

References

[1] Kosko, B., "Bidirectional associative memory", *IEEE Trans. System Man Cybernet*, **18**(1), 49–60, Jan./Feb., 1988.

[2] Wang, C.-C. and Don, H.-S., "An analysis of high-capacity discrete exponential BAM", *IEEE Trans. on Neural Networks*, **6**(2), 492–496, Mar., 1995.

[3] Krambeck, R. H., Lee, C. M. and Law, H.-S., "High-speed compact circuits with CMOS", *IEEE J. Solid-State Circuits*, **17**, 614–619, June, 1982.

[4] Goncalves, N. F. and De Man, H. J., "NORA: A race-free dynamic CMOS technology for pipelined logic structures", *IEEE J. on Solid-State Circuits*, **18**, 261–266, June, 1983.

[5] Gu, R. X. and Elmasry, M. I., "All-N-logic high-speed true-single-phase dynamic CMOS logic", *IEEE J. on Solid-State Circuits*, **31**(2), 221–229, Feb., 1996.

[6] Afghahi, M., "A robust single phase clocking for low power high-speed VLSI application", *IEEE J. of Solid-State Circuits*, **31**(2), 247–253, Feb., 1996.

[7] Yuan, J. and Svensson, C., "High-speed CMOS circuit technique", *IEEE J. on Solid-State Circuits*, **24**, 62–70, Feb., 1989.

[8] Lee, C. M. and Szeto, E. W., "Zipper CMOS", *IEEE Circuits Devices Mag*., pp. 10–16, May, 1986.

[9] Zhang, D. and Elmasry, M. I., "VLSI compressor design with applications to digital neural networks", *IEEE Trans. on VLSI Systems*, **5**(2), 230–233, June, 1997.

Authors' Biographies

Chua-Chin Wang was born in Taiwan, in 1962. He received the B.S. degree in electrical engineering from National Taiwan University, Taiwan, in 1984 and the M.S. and Ph.D. degrees in electrical engineering from State University of New York, Stony Brook, in 1988 and 1992, respectively. Currently he is a Professor in the Department of Electrical Engineering, National Sun Yat-Sen University, Taiwan. His research interests include low-power logic and circuit design, VLSI design, and neural networks and implementations.

Chenn-Jung Huang was born in Hualien, Taiwan, in 1961. He received the B.S. degree in electrical engineering from National Taiwan University, Taiwan, in 1984 and the M.S. degree in computer science from University of Southern California, Los Angeles, in 1987. He is currently completing requirements for the Ph.D. degree in electrical engineering at National Sun Yat-Sen University, Taiwan. His current research interests are computer arithmetic, computer communication networks, and neural networks.

Po-Ming Lee was born in Tainan, Taiwan, in 1973. He received the B.S. degree in computer science and engineering from Yuan Zu University, Taiwan, in 1995 and the M.S. degree in electrical engineering from National Sun Yat-Sen University, Taiwan, in 1999. He is currently working toward the Ph.D. degree in electrical engineering at National Sun Yat-Sen University. His current research interests are VLSI design and computer arithmetic.