A COMPARISON OF TWO ALTERNATIVE ARCHITECTURES OF DIGITAL RATIOED COMPRESSOR DESIGN FOR INNER PRODUCT PROCESSING

C.-C. Wang, C.-J. Huang, & P.-M. Lee

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

ABSTRACT

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 novel architectures to arrange digital ratioed compressors are proposed 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.

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 $Y = th\left(\sum_{i=1}^{n} Y_i \cdot \left(X_i \cdot X\right)\right)$, 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.

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 forms of the digital ratioed compressors building blocks based on the *3-2* compressors are proposed, where the problems mentioned above are all resolved. An analytical form of carry propagation delay estimation for these novel 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:

$$S = (a \oplus c)\, b' + (a \oplus c)'\, b = F\, b' + F'\, b$$
$$C = (a \oplus c)\, b + (a \oplus c)'\, c = F\, b + F'\, c \qquad (1)$$

where F denotes $(a \oplus c)$. Then, a typical *3-2* compressor is shown in Fig. 1. The feature of such a compressor is that the output represents the number of 1's given in inputs.

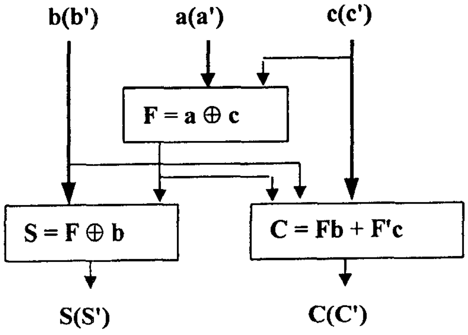
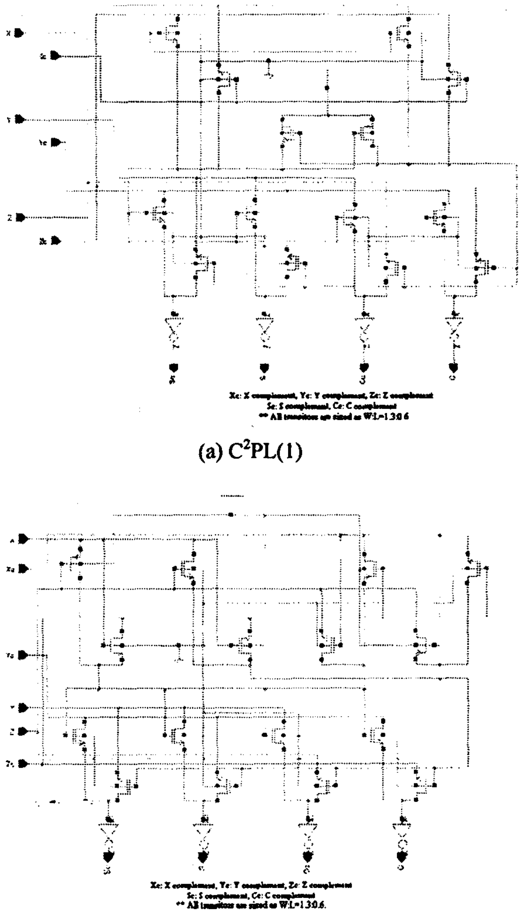


Fig. 1. A *3-2* compressor building block.

0-7803-5471-0/99/$10.00©1999 IEEE

2.2 Ratioed 3-2 compressor design

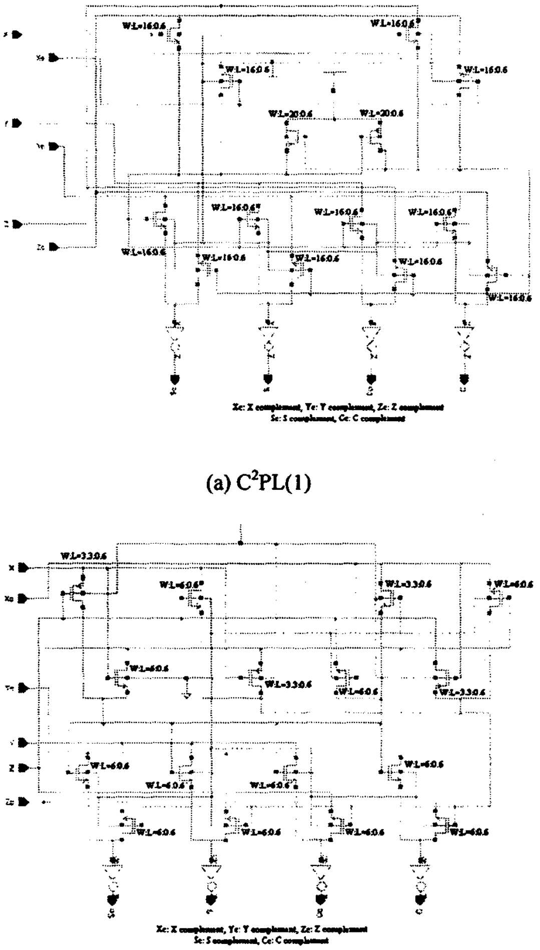
Though a *3-2* compressor can be realized by a full adder, and Zhang *et al.* [9] proposed a C^{2}PL design for *3-2* and *7-3* compressors, several design issues as addressed in Section 1 are still ignored in their work. Fig. 2 shows the schematic diagrams for the two types of *3-2* compressors based on complex complementary pass-transistor logic (C^{2}PL) 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 Fig. 3. 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.


Xc: X complement, Yc: Y complement, Zc: Z complement
Sc: S complement, Cc: C complement
** All transistors are sized as W/L=1.3/0.6

(a) C^{2}PL(1)

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

Fig. 2. Schematic diagram for C^{2}PL *3-2* compressor in original design.


Xc: X complement, Yc: Y complement, Zc: Z complement
Sc: S complement, Cc: C complement

(a) C^{2}PL(1)

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

Fig. 3. Schematic diagram for re-designed C^{2}PL *3-2* compressor.

2.3 The first general form of ratioed compressor building blocks

A *7-3* compressor building block can be constructed by cascading four *3-2* compressors as shown in Fig. 4. A *15-4* compressor building block can also be formed similarly with two *7-3* compressors and two *3-2* compressors, as shown in Fig. 5. Basing on this design methodology, a general form for a *(*2^n*-1)-n* compressor building block is composed of two *(*2^{n-1}*-1)-(n-1)* compressors and *(n – 1) 3-2* compressors.

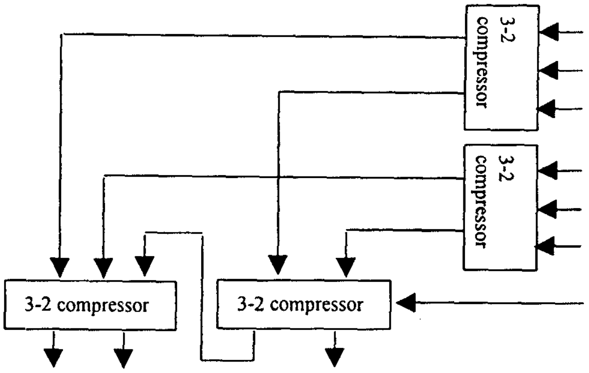


Fig. 4. A *7-3* compressor building block.

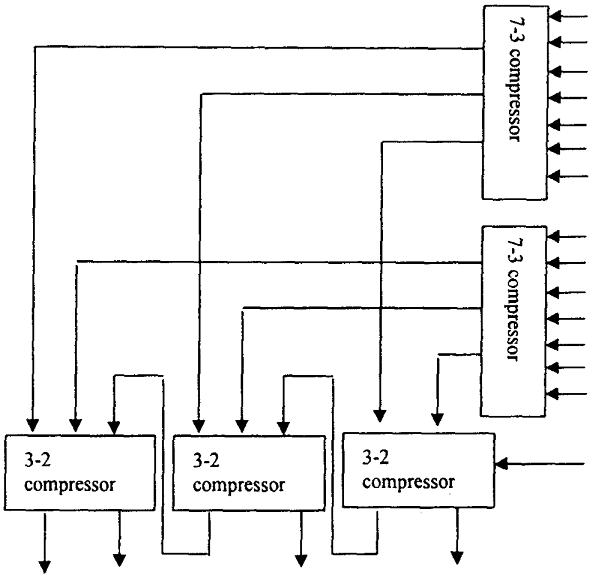


Fig. 5. A *15-4* compressor building block for General Form I.

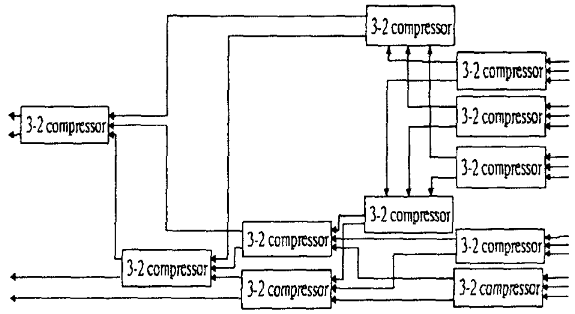


Fig. 6 A *15-4* compressor building block for General Form II.

2.3.1 Carry propagation delay equations

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 $2^n\text{-}1$ bits are applied on the $(2^n\text{-}1)\text{-}n$ compressor block. By observing the structure of the compressor blocks, we can deduce D_2, D_3, and D_n as follows:

$$D_2 = 1$$

$$D_3 = 1 = 1 + 2 = 2 + D_2$$

$$D_n = n - 1 + D_{n-1}, \qquad n \geq 3. \tag{2}$$

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 $\dfrac{\ln\frac{n}{M}}{\ln\frac{n}{2^n-1}}$, 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 applied on the $(2^n\text{-}1)\text{-}n$ compressor building blocks can be shown as follows:

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

2.4 The second general form of ratioed compressor building blocks

The second form we propose to improve the carry propagation delay of the critical paths is shown in Fig. 6. This architecture, inspired by the design methodology of systolic arrays, consists of parallelized *3-2* compressor building blocks only at every processing stage. The total delays of $(2^n\text{-}1)\text{-}n$ compressors, where n from 2 to 17, have been computed by a program and can be formulized as follows:

$$D_n = \begin{cases} 2n-3 & 2 \leq n \leq 6 \\ 2n-4 & 7 \leq n \leq 17 \end{cases} \tag{5}$$

In order to realize the performance improvement of this general form in the carry propagation delay of the critical paths, the following table shows the comparison with the other architecture we presented above.

Form \ *n*	2	3	4	5	6	7	8	9	10	11
I	1	3	6	10	15	21	28	36	45	55
II	1	3	5	7	9	10	12	14	16	18

Table 1: Total delay comparison of two general forms

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 follows:

delay \ circuits	The original *3-2* compressor				The re-designed *3-2* compressor			
	C^2PL(1)		C^2PL(2)		C^2PL(1)		C^2PL(2)	
	carry	sum	carry	sum	carry	sum	carry	sum
rise delay(ns)	0.26	0.31	0.42	0.35	0.32	0.36	0.41	0.34
fall delay(ns)	0.87	0.83	0.87	0.87	0.24	0.43	0.39	0.42

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

3.2 Delay simulations

The Verilog simulations are performed *20000* iterations for the first general form and the systolic-array form of *127-7* compressor building blocks, respectively. Table 3 illustrates the comparison of carry propagation delay for the two architectures of *127-7* compressor building blocks when they are applied in *127* data inputs summation.

delay \ circuits	Form I		Form II	
delay	C^2PL(1)	C^2PL(2)	C^2PL(1)	C^2PL(2)
delay (ns)	10	9	4	5

Table 3: The comparison of carry propagation delay for the two architectures of *127-7* compressor building blocks

The results demonstrate that the systolic-array form compressors indeed lead the least carry propagation delay.

4. CONCLUSION

In this paper we have proposed two novel organizations of basic compressor building blocks which can be adopted in the implementation of digital neural networks. The 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 are derived and used for exploration of the superior architecture of ratioed compressor building blocks for the digital neural network applications. Our simulation results show that the systolic-array architecture gives a sub-optimal performance through the parallelized arrangement of *3-2* compressors at each stage of processing.

5. REFERENCES

[1] B. Kosko, "Bidirectional associative memory," *IEEE Trans. System Man Cybernet*, vol. 18, no. 1, pp.49-60, Jan./Feb. 1988.

[2] C.-C. Wang, and H.-S. Don, "An analysis of high-capacity discrete exponential BAM," *IEEE Trans. on Neural Networks*, vol. 6, no. 2, pp. 492-496, Mar. 1995.

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

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

[5] R. X. Gu, and M. I. Elmasry, "All-N-logic high-speed true-single-phase dynamic CMOS logic," *IEEE J. on Solid-State Circuits*, vol. 31, no. 2, pp. 221-229, Feb. 1996.

[6] M. Afghahi, "A robust single phase clocking for low power high-speed VLSI application," *IEEE J. of Solid-State Circuits*, vol. 31, no. 2, pp. 247-253, Feb. 1996.

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

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

[9] D. Zhang, and M. I. Elmasry, "VLSI compressor design with applications to digital neural networks," *IEEE Trans. on VLSI Systems*, vol. 5, no. 2, pp. 230-233, June 1997