International Journal of Fuzzy Logic and Intelligent Systems, vol. 4, no. 1, June 2004 pp. 79-83

Low Power Design of the Neuroprocessor

A. S. Pandya*#, Ankur Agarwal* and G. Y. Chae**

***Dept. of Computer Science & Engineering, Florida Atlantic University,**
Boca Raton, FL-33431, USA
 # Div of Information and Computer Engineering
**** IT Design Research Center**
Silla University, Busan 617-736, Korea

Abstract

This paper presents the performance analysis for CPL based design of a Low power digital neuroprocessor. We have verified the functionality of the components at the high level using Verilog and carried out the simulations in Silos. The components of the proposed digital neuroprocessor have also been verified at the layout level in LASI. The layouts have then been simulated and analyzed in Winspice for their timing characteristics. The result shows that the proposed digital neuroprocessor consistently consumes less power than other designs of the same function. It can also be seen that the proposed functions have lesser propagation delay and thus higher speed compared to the other designs.

Key word : CPL, Low power, digital neuroprocessor, VLSI

I. Introduction

There has been a substantial increase in the programmable hardware and the usability of Hardware Design languages in the last decade. This has led to the evolution of designs in hardware which were traditionally done in System on chip (SOC). Despite many years of studies in neural networks, it is only due to advancement in programmable hardware that actual implementation for study and there applications have become practical. Though there are many examples of neural network codes running on von Neumann computers, there are still few commercially available neural networks implemented in hardware.

Artificial neural networks over last two decades have evolved to combine a wide variety of training platforms and architectures. However biased results have been produced by features of serial processing, as most of the studies are done through computer simulations. Several previous researchers in the field of digital neuroprocessors have reported various application specific integrated circuit (ASIC) implementations of neuroprocessors in high speed Complementary Metal Oxide Semiconductors (CMOS) and very large scale integration (VLSI) [1] [2] [3] [4] [5]. Such efforts could allow the researchers to exploit the remarkable parallel and distributive characteristics of neurocomputing.

A very common type of neural network is the feed-forward Multi-layer perceptron (MLP) with two or more layers as shown in Figure 1. In this architecture, the processed result of one neuron is sent to only those neurons, which are in the subsequent layer. Due to this various algorithms for information processing in the neuron have been proposed and implemented in the core of the neuroprocessors [2]. Such networks can be trained by algorithms such as, the well known back-propagation method.

Neuroprocessor architectures are being proposed as a plausible approval for diverse problems in complex areas of mathematics, modeling of information processes in the nervous systems, implementation of information converters, realization of neuroprocessor associative memory and building of learning recognition systems.

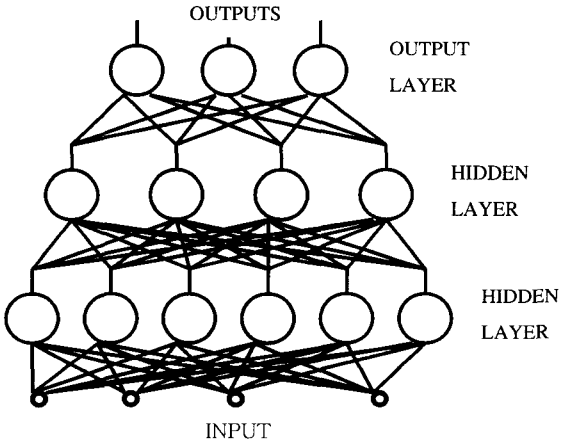


Fig. 1. Multilayer perceptron Architecture

Figure 2 shows the basic implementation for the hardware of a neuroprocessor, which can be seen as a single neuron which has all inputs feeding into a summing junction whose output is fed into a hard limiting activation function. We can then build a multilayer perceptron of Figure-1 by replacing each neuron with such a low-power neuroprocessor.

Manuscrip received Feb. 10, 2004; revised Apr. 2, 2004.
This research was partly supported by a grant from Korean Ministry of Information and Communication.

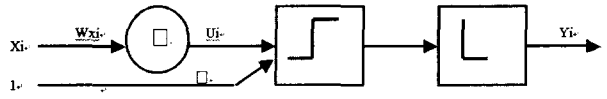


Fig. 2. Basic Hardware Implementation of the Neuroprocessor

The model presented here is the detailed explanation of the basic architecture of a neuroprocessor. All the inputs along with the single input of the neuron are fed into a summing junction, where the summing of all the inputs is carried out by adding up the weights and changing the sign of a weight if the input value is negative. The output from this junction is extracted to be fed to a hard limiting activation function. The activation function takes the sign of the resultant summation and presents it to the latch. The final result is held in the latch. The latch passes the value on to the output on a positive clock edge.

In the basic architecture [3] the input values are not in the real domain. However for a better and efficient output the inputs should have real values and the activation function should be sigmoid or hyperbolic instead of threshold. Thus new architectures are being proposed for the same. Some the few proposed architectures include the bit serial and the bit parallel architectures used for the hardware of a neuroprocessor. We shall first analyze the bit serial architecture followed by the bit parallel before we propose our novel architecture for the neuroprocessor.

One design which exploits the property of parallelism is a NeuroMatrix NM6403 neuroprocessor [4]. This device takes full advantage of the neural network concept and provides powerful computational capabilities, including its massive parallelism, redundancy and robustness. An important feature of this processor apart from parallelism is the presence of a static and dynamic memory having a wide range of time parameters without an external controller which is due to the availability of programmable memory interface units. Parallelism may seem sufficient for the fast evaluation of the large networks however it results in employment of large hardware circuitry which reduces its cost effectiveness.

An alternative to this is compact bit serial architecture [5], which is cost efficient. In this design the same neurons are reused in successive layers of the network thereby reducing the total number of neurons from those used in the largest layer. This requires less interconnections and hardware which leads to better efficiency in terms of power consumption and cost effectiveness. This also causes iterative usage of the same circuitry in a time multiplying scheme to implement successive layers in the neural network. However this process slows down the computation.

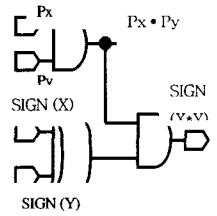
In this paper we have proposed a novel design of the neuroprocessor based on addition and multiplication operations which makeup the summation unit in Figure 2. The main aspect of this design is the low power circuitry we are using for its implementation. Here we are proposing the design for the summation unit of the neuroprocessor such that it employs less hardware than the other similar architectures. With this we are able to exploit the property of the parallelism by replication of the basic operation. The new design of neuroprocessor has been employed by using complementary pass transistors. By looking into aspects like power consumption, cost effectiveness, speed and wiring complexity we shall analyze the new design. These results are obtained from Silos, LASI and Winspice simulations of the proposed designs.

The interesting fact in design of proposed neuroprocessor is the design of the arithmetic operations, which shows a significant reduction in the number of nMOS and pMOS transistors employed in the design. This results in increasing the speed and a significant reduction in the power consumption and the layout complexity.

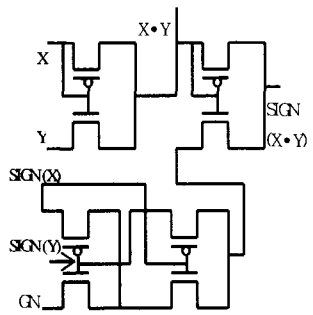
2. Components of Proposed Design of Neuroprocessor

2.1 Novel Design of Multiplication Circuit

The block diagram of the standard and the proposed design of the multiplication circuit are shown in Figure 3 (a) and (b) respectively. The proposed design of the multiplication circuit is based on the transfer gate technology [6]. Figure 4 shows the layout analysis of the proposed design of the low power multiplication circuit using LASI. Extensive analysis demonstrates that the proposed design of the low power multiplication circuit behaves as the standard CMOS design for the same as far as the logical function is concerned.



(a)



(b)

Fig. 3 (a) and (b). Block Diagram for the Standard Design and the Proposed Low Power Design of the Signed Multiplication Circuit

This can also be seen from the Verilog gate level coding style of both circuits as well. Extensive analysis of the layout of the proposed multiplication circuit shows that, the number of transistors employed for its implementation is drastically

reduced as compared to its standard design. In the conventional design, the multiplication circuit is achieved by the use of twenty transistors (pMOS and nMOS), while in the proposed case eight transistors are used for the same design. This can be seen from the layout analysis in Figure 4. This significant reduction in the transistor count increases the speed of the circuit and reduces the switching frequency which is an important parameter contributing towards the reduction in the Dynamic Power consumption. Another aspect to note in the proposed design is the absence of the explicit VDD supply requirement in the proposed design of the multiplier circuit.

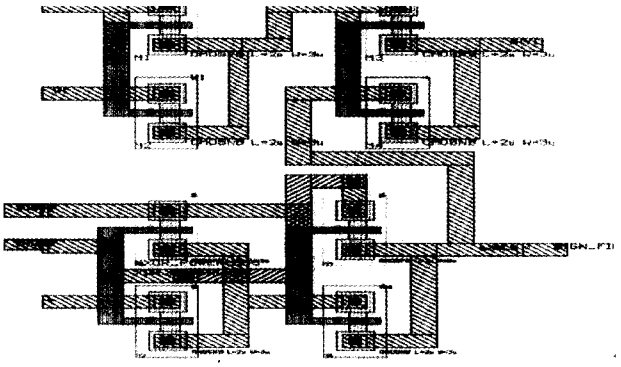


Fig. 4. Layout Shown for the Proposed Design of the Low Power Multiplier Circuit Using LASI

Winspice simulation results for the layout analysis of proposed design of multiplication circuit are shown in Figures 5 and 6.

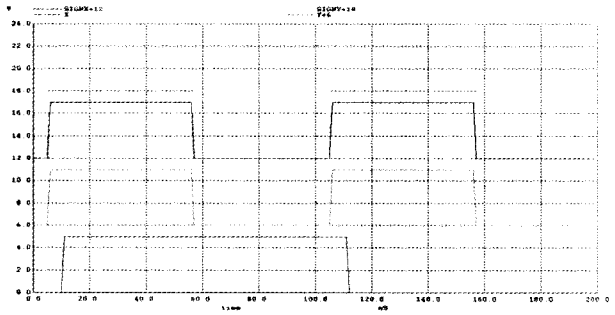


Fig. 5. Winspice Simulation Result of the LASI Layout for the Proposed Design Multiplier Circuit

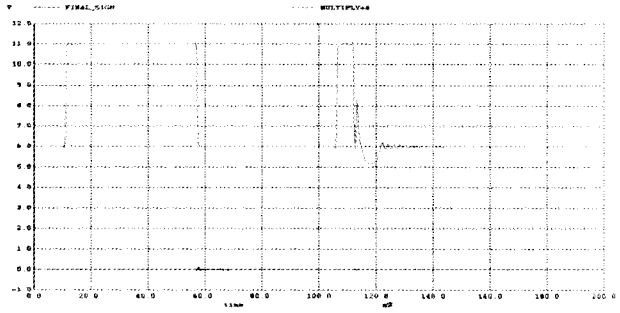


Fig. 6. Winspice Simulation Result of the LASI Layout for the Proposed Design Multiplier Circuit

2.2 Low Power Design of the Addition Circuit

A standard CMOS design of the Full Adder circuit employs 28 transistors. A low power design for the Full Adder was proposed by Al-Sheraidah [7]. This design uses only twelve gates against twenty-eight as in the standard design, shown in Figure 7. This design is based on the transfer gate, which uses low power due to the absence of an explicit power supply. There are some trade-offs in this design in terms of the voltage swing and the power consumption. Due to the absence of explicit VDD and GND this circuit consistently consumes less power than the standard design of the adder circuit.

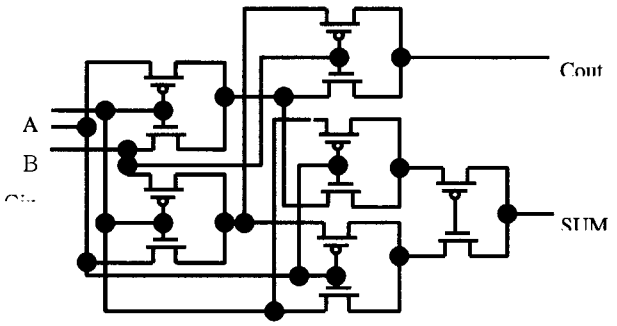


Fig. 7. Block Diagram for the Low Power Design of the Full Adder Circuit

3. Analysis of Power Consumed in the Proposed Designs

The main reason for implementing the majority of contemporary high-complexity designs in static CMOS is the almost complete absence of power consumption in steady-state mode. According to equation (1) [8] [9], the average power dissipation P_v in digital CMOS circuits includes three distinct components.

$$P_v = P_{short\ circuit} + P_{switching} + P_{leakage} + P_{static} \quad (1)$$

It can be seen from the above equation that the dynamic power dissipation for digital CMOS circuits depends upon clock frequency, transition activity, node capacitance, short circuit current and the power supply V_{DD}.

3.1 Dynamic consumption due to load capacitance

Nodes in the digital circuits toggle between the two logic states, 0 & 1. During the transition from one state to another, node capacitances need to be charged & discharged. The current passing through either p-channel or n-channel transistors while charging or discharging node capacitances causes the capacitive component Pcap of the total power consumed. P_{cap} is given by equation (2).

$$P_{switching} = P_{cap} = \alpha \cdot C_L \cdot V_{DD}^2 \cdot f_{clk} \quad (2)$$

Here f_{clk} is the clock frequency, V_{DD} is frequency supply voltage, C_L is the load capacitance and alpha is the transition activity. It can be concluded that as the transition of the states

increases the power dissipation increases. In the proposed model due to the reduced number of transistor count the transition activity is reduced by 60% resulting in reduced power consumption.

The capacitive load Pcap, in equation (2) originates from the capacitance between the gate and diffusion and the interconnecting metal and Polysilicon layers in our drawn layouts. This can be substantially reduced by employing fewer transistors in our design and reducing the size of the transistor i.e. the length and the width of the channel to the minimum possible size. As mentioned above in equation (2), there is the square law dependence of capacitive switching power on the supply voltage, V_{DD}. Therefore, reducing the supply voltage is an effective means for reduction of Pcap. In many cases, supply voltage reduction and speed-up design techniques go along with reduction of the clock frequency f_{clk} which reduces the capacitive switching power even further.

3.2 Leakage Power

Ideally, digital CMOS circuits should not exhibit any static power consumption. However, due to the non ideal sub threshold behavior of MOSFETs, there is a leakage current Ileak flowing from the positive power supply to the ground even in the static case resulting in the leakage power Pleak, which is given by the equation (3)

$$P_{leak} = I_{leak} \cdot V_{DD} \quad (3)$$

Here, reduction in the requirement for the supply voltage and the number of transistors results in a significant reduction of power consumption.

3.3 Short Circuit power

Short circuit power (P_{short}), is expressed as follows:

$$P_{short\ circuit} = (1/2) \cdot \alpha \cdot (t_r \cdot I_{short,max,r} + t_f \cdot I_{short,max,f}) \cdot V_{DD} \cdot f_{clk} \quad (4)$$

Where, $I_{short,max,r}$ is the peak of the current flowing from positive power supply to ground when n- and p-channel transistors are conducting simultaneously for an infinitely small moment during node transition and tr/f is the fall and rise time of the node voltage. This short circuit power decreases with the decreasing switching activity alpha and decreasing clock frequency 'f_{clk}'.

However clock frequency is usually regarded constant in order to fulfill some architectural requirements. Hence reduction in short circuit power can be achieved by reducing the number of transistors and thereby reduction in the switching activity. It can be observed that in all the cases (equations 2-4) the power consumed directly depends upon V_{DD} i.e. the supply voltage. Due to the absence of the explicit V_{DD} and GND supply in our circuit (see Figures 3b and 7) theoretically the dynamic power consumed will be negligible, whereas static power consumption will be responsible for the total power consumed by the device.

4. Conclusion

According to the equations 1, 2, 3 and 4, power consumption mainly depends upon V_{DD}, number of transistors and switching frequency. As a result, when there is no explicit V_{DD} connection, as is the case in our circuit designs, the short circuit power consumption and the leakage power consumption are reduced substantially. Of course, there will be static power consumption, but switching power consumption is negligible. It can also be seen from the proposed design of the neuroprocessor that it employs twenty transistors in its design against forty eight transistors as in the standard design. This saves of 58% of transistors. This saving of 58% transistors accounts for about 35% of the saving in the power requirements. Now due to the absence of explicit V_{DD} and GND supply requirement the total saving of the power comes to 60%. This makes the proposed designs an ideal one for digital neuroprocessors, which require low power.

However, due the absence of explicit VDD or GND connection the pull-up network and the pull-down network in the circuit are never formed as in the conventional CMOS technology [9].

References

[1] Hammerrstron D.: A VLSI Architecture for High-performance, Low Cost, On-chip Learning, IEEE International joint Conference on Neural Networks, pp. 537-544, 1990.

[2] Pandya A. S., Macy R. B.: Pattern Recognition with Neural Networks in C++. IEEE Press, CRC Press, Boca Raton, FL 1995.

[3] Chapman R., Sutankayo S.: Implementation of Neural Network Designs: Rochester Institute of Technology Final Report 1997.

[4] Chevtchenko P.A., Fomine D. V. Tcherikov V. M., Vixne P.E:: Neuroprocessor NeuroMatrix NM6403 Architecture Overview, with Application to Signal Processing, SPIE Proceedings, Ninth Workshop on Virtual Intelligence/Dynamic Neural Networks, Vol. 3728, 1991, pp. 253-264.

[5] Tawer R.: Compact Bit-Serial VLSI Neuroprocessor for Automativ Use, Nasas Jet Propulsion Laboratory, Techinical Report 1996.

[6] Agarwal A.: Low Power Design of an ALU, MS Thesis, Florida Atlantic University, 2003.

[7] Al-Sheraidah A.: Novel Multiplexer Based Architectures for Full-Adder Design, MS Thesis, Florida Atlantic University 2000.

[8] Rabaey J. M.: *Digital Integrated Circuits, A Design Perspective*, Prentice Hall, 1995.

[9] Chandrakasan A.P., Sheng S. Brodersen R.W., Low-Power CMOS Digital Design, IEEE Journal of Solid State Circuits, Vol. 27, No. 4, pp 473-483, April 1992.

Dr. A.S. Pandya is a professor at the Computer Science and Engineering Department, Florida Atlantic University. He has published over 100 papers and book chapters, and a number of books in the areas of neural networks and ATM networks. This includes a text published by CRC Press and IEEE Press entitled "Pattern Recognition using Neural Networks in C++". He consults for several industries including IBM, Motorola, Coulter industries and the U.S. Patent Office. He received his undergraduate education at the Indian Institute ofTechnology, Bombay. He earned his M.S. and Ph.D. in Computer Science from the Syracuse University, New York. He has worked as a visiting Professor in various countries including Japan, Korea, India, etc. His areas of research include VLSI implementable algorithms, Applications of AI and Image analysis in Medicine, Financial Forecasting using Neural Networks.

Ankur Agarwal is a PhD. student in Computer Engineering department at Florida Atlantic University, Boca Raton, FL. He pursued his MS at Florida Atlantic University and BE from Pune University, India. He also holds two post graduate diplomas in VLSI design and in Embedded System Design. He has published several papers in VLSI design area and in computer architecture. Currently he is involved in research work related to Embedded System Design and Hardware-Software Co-design.

Gyoo-Yong Chae
Received his M. S. and the Ph.D. degrees in Department of Industrial Engineering from Konkook University, Seoul, Korea. Present, he is a Research Committee of IT Design Research Center at the Silla University. His research interests include Fuzzy Logic, Data Mining and Data Base.