Pulse Coded Neural Network Implementation In VLSI

Nasir Shaikh-Husin
Fakulti Kejuruteraan Elektrik
Universiti Teknologi Malaysia
81310 UTM Skudai
Malaysia
e-mail: nasir_s_h@yahoo.com

Chang Wooi Po
Agilent Technologies (Malaysia) Sdn. Bhd.
Bayan Lepas FTZ
11900 Penang
Malaysia
e-mail: wooi-po_chang@agilent.com

Abstract: A neural network that encodes signals in terms of pulses has been designed and fabricated. The neural network components are described in detail. As a test case, a two-layer network is implemented. A preliminary test result shows some promise and some limitations of the design.

Keywords
Pulse modulation, neural network, MOS integrated circuit.

I. INTRODUCTION

Artificial neural network refers to processors that at some level reflect what is known about the structure of the brain. It is a parallel computation technique inspired by biological process. Although there are many neural network paradigms, most models share common characteristics of parallelism in computation performed by relatively simple processors. No single processor performs a critical role or stores crucial information. The power of computation is derived from the collective cooperation of those simple processors. The cohesion in computation among processors makes them rather robust to noise. This property gives neural networks the ability to produce appropriate response to noisy or incompletely specified inputs.

Several abstract neural network models have been introduced. The pulse coded neural network (PCNN) models are a compromise between abstract processing units and detailed biological neural models. The most distinguishing feature of pulse coded models is that their inputs and outputs are in the form of pulses [1]-[3]. All communications among processing units (called neurons) and the environment occurs in the form of binary voltage pulses. Pulse coded techniques combine both digital and analog processing. The use of pulses for communication has several advantages in terms of hardware implementation. Pulses can be communicated on a single wire just like an analog signal, but with noise immunity and ease of transmission of a digital signal [4]. The pulses can be generated and detected using simple circuits. Design procedure for the circuits is based on digital techniques, producing reliable and robust circuits [5].

There are several methods that can be used to encode analog information in terms of pulses [1]. Among the modulation schemes are pulse amplitude (PAM,) pulse width (PWM,) and pulse frequency (PFM.) In this paper, a modulation scheme similar to PFM is used. It is called pulse density modulation [5]. Instead of pulses with variable frequency, the density i.e. the number of pulses per unit time increases for increasing analog signal. The amplitude and pulse width of individual pulses are fixed. An example of how an analog signal is encoded is shown in Fig. 1.

A PCNN chip was designed and fabricated using CMOS 2-μm p-well process. The components of the neural network are described in Section II. Section III describes how the PCNN chip operates. Some preliminary test results from the fabricated prototype are presented in Section IV.

II. PCNN COMPONENTS

A. Self-resetting Neuron

A simplified circuit for the self-resetting neuron is shown in Fig. 2(a) [5]-[6]. The circuit consists of an integrating capacitor C_{sum}, discharge resistor R, two switches S1 and S2, and a Schmitt trigger.

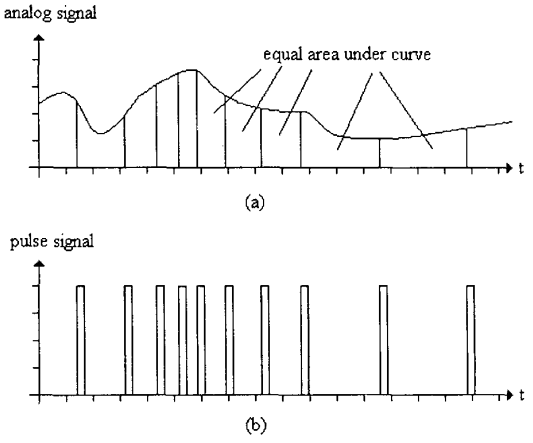

Fig. 1. Pulse density modulation. (a) Analog signal. (b) Encoded pulses.

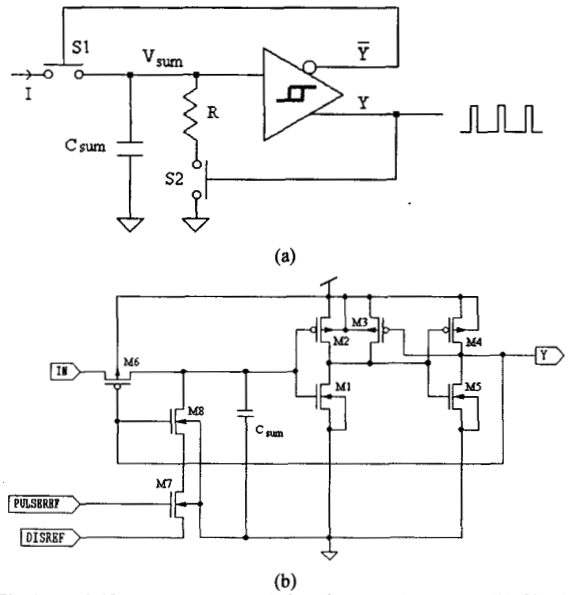

Fig. 2. Self-resetting neuron. (a) Functional block diagram. (b) Circuit schematic.

To see how the neuron operates, we assume the capacitor is initially not charged and hence V_{sum}, representing the instantaneous membrane potential of a natural neuron, is LOW. This causes the Schmitt trigger output Y to be LOW and $\overline{Y}$ to be HIGH. These two outputs control the two switches and thus S1 is now closed and S2 is open. All the synapses that are connected to the neuron input now can charge the capacitor in the form of charge or current pulses. The input currents are integrated on the capacitor until the capacitor voltage V_{sum} exceeds the upper threshold voltage of the Schmitt trigger V_{th}. At this point, Y rapidly switches to HIGH and $\overline{Y}$ becomes LOW. Consequently, switch S1 is now open and S2 is closed. The neuron now ignores all inputs from the synapses. At the same time, the capacitor is discharged through R until the voltage on the capacitor breaches the lower threshold voltage of the Schmitt trigger V_{tl}. When this happens, each Schmitt trigger output toggles again, thus changing the position of the switches and the integration/discharge cycle repeats.

Fig. 2(b) shows the circuit schematic of the neuron circuit [6]. Transistors M1-M5 form the Schmitt trigger. Switches S1 and S2 are represented by pMOS transistor M6 and nMOS transistor M8 respectively while M7 corresponds to the resistance R. The source of M7 can be grounded. However, to increase flexibility of the circuit, the source is connected to DISREF. This signal can then be connected to ground or to a bias voltage. Another external bias voltage, PULSEREF, controls the conductance of M7. A high value of PULSEREF will result in a relatively large conductance, or in other words a small resistance. This will result in small pulse duration at the output of the Schmitt trigger because V_{sum} is discharged quickly. It takes only a short time for the Schmitt trigger input to reach the lower threshold voltage from the upper threshold. The inter-pulse duration, that is the period where the output is LOW, is

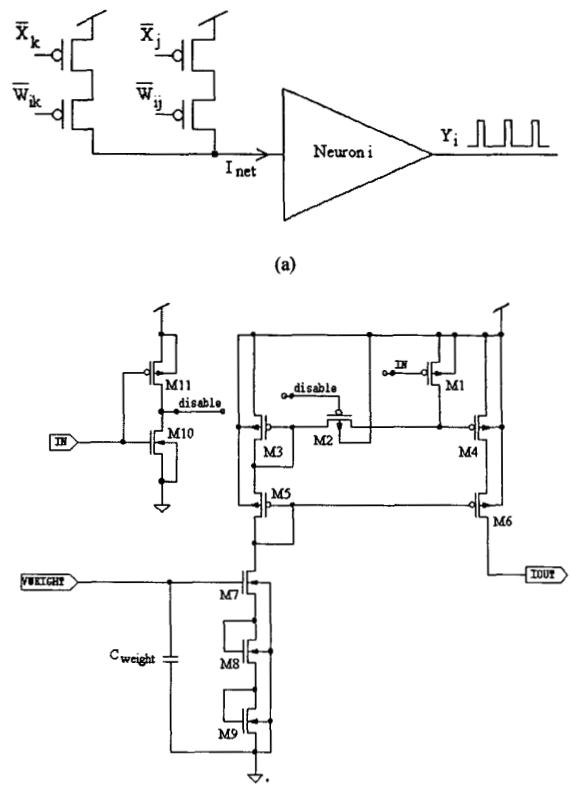

Fig. 3. Synapse circuit. (a) Synapse operation. (b) Circuit schematic.

governed by the size of C_{sum} and the amount of input current.

B. Synapse Circuit

The output of a neuron is transmitted to other neurons via synapses. Voltage pulses received by a synapse from a neuron will be converted to current pulses, which then serve as input to the next neuron. The synapse operation is depicted in Fig. 3(a). A neuron is shown to be connected to two synapses. Each synapse has two inputs: $\overline{X}$ and $\overline{W_i}$. The $\overline{X}$ input is connected to a neuron output while input $\overline{W_i}$ is the weight input, which regulates the amount of current that flows from a synapse when there is a voltage pulse. This weight input represents the connection strength, or synaptic weight, between a particular neuron to another. Currents from all synapses connected to a neuron are summed at the input. The net current modulates the pulse train emanating from the neuron.

A realistic circuit for the conceptual synapse of Fig. 3(a) is shown in Fig. 3(b). Ignoring M1 and M2 for now, M3-M6 form a cascode current-mirror. This current mirror has a high output resistance, hence the current sourced into the neuron is less susceptible to the value of the voltage at the neuron input. M1 and M2 determine whether the current mirror can function or not. If input IN is LOW, i.e. no

pulse, M2 is off and M1 shorts the gate of M4 to VDD. This condition prevents the current mirror from functioning. However, if IN is HIGH (pulse is present,) M1 is off and M2 is on, and the current mirror is operational. Hence, the input signal IN determines whether there is any current coming from that synaptic connection. This signal comes from the output of another neuron.

Whenever there is a pulse, the amount of synaptic current is controlled by the current sink formed by M7-M9. This current sink in turn depends on the bias voltage VWEIGHT that is stored by the capacitor C_{weight}. Each synapse must have its own VWEIGHT to represent the connection strength between a neuron to another. For the current sink, instead of utilizing a single transistor, three transistors are stacked together. This scheme enables the transistors to operate in subthreshold region, resulting in small current at the output of the synapse. This not only reduces power dissipation of the synapse, but it also enables the integrating capacitor at the neuron input, C_{sum}, to be small. Stacking the transistors also increases the dynamic range of VWEIGHT input while still keeping circuit operation in subthreshold.

III. NON-VOLATILE SYNAPTIC WEIGHT STORAGE

Since the PCNN chip uses a capacitor to store synaptic weight, a refresh mechanism must be instituted in order for the weight information to be retained. Hence, each synapse has its own VWEIGHT controller. The non-volatile weight storage technique employed for this chip requires that an external staircase refresh voltage be globally connected to all controllers. The increasing staircase voltage is continuously compared to the current VWEIGHT. When the refresh voltage surpasses VWEIGHT, a transistor switch connects the refresh input to the weight capacitor. The refresh operation takes place simultaneously for all synapses. Provided that VWEIGHT does not drop too much (less than staircase increment,) the refresh voltage restores VWEIGHT to its initial value.

Apart from the refresh function, VWEIGHT controller can also write a new weight value to the weight capacitor. This function is important because usually a neural network needs to adapt its response to different classes of input data. Lastly, the controller can be programmed to read the stored weight for monitoring purposes. Fig. 4 shows a high level schematic for VWEIGHT controller [6]. The controller function can now be explained in detail.

The global staircase VREFRESH input is continuously compared to the instantaneous weight voltage VWEIGHT. When the former exceeds the latter, the comparator output becomes HIGH, which in turn activates the output of the monostable multivibrator. As a result, transistor M4 turns on and connects VREFRESH to VWEIGHT. A dummy transistor, M5, is included to help alleviate switch feedthrough effect when M4 is turned on by monostable

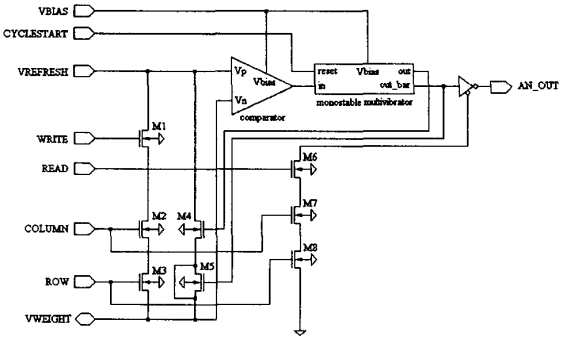

Fig. 4. Weight refresh circuit.

output. The duration of M4 turned on can be controlled by bias voltage VBIAS.

For read or write operation, it must be done to a single synapse one at a time. To overwrite (or initialize) VWEIGHT of a synapse, the ROW and COLUMN inputs must be activated. The WRITE input then must wait until VREFRESH reaches the desired value. During this time, and only during this time, WRITE must be asserted. Transistors M1-M3 are now turned on and the instantaneous value of VREFRESH is transferred to VWEIGHT. To read the value of VWEIGHT, signals ROW, COLUMN, and READ are asserted. The tristate inverter is now enabled. An output signal AN_OUT will be asserted precisely at the moment VREFRESH first exceeds VWEIGHT. By decoding the position of AN_OUT pulse in VREFRESH staircase signal, the value of VWEIGHT can be determined.

The comparator circuit used in the controller is a simple, seven-transistor circuit [7]. The monostable circuit is shown in Fig. 5. It consists mainly of a Schmitt trigger and two transistor switches that control the charging and discharging of parasitic capacitance at the trigger input. The reset signal must be asserted, in the form of a short pulse, before the trigger input (in) can be active. The reset pulse is needed to turn on M7. The output now will wait the trigger input to come. When this pulse is active, the output becomes HIGH and causes M9 and M6 to turn on. M7 now turns off while Vbias determines the duration of the output pulse by controlling the rate of capacitor discharge.

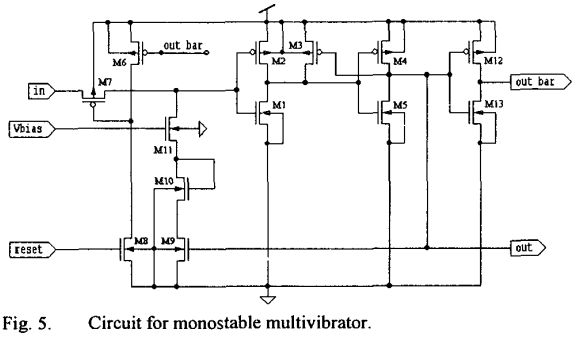

Fig. 5. Circuit for monostable multivibrator.

IV. PROTOTYPE NEURAL NETWORK

A. Fabricated IC

As a prototype, a two-layer PCNN has been fabricated in 2 μm p-well CMOS process with two metal layers and two poly layers. The chip was fabricated through MOSIS service. The poly layers are used to create the capacitors, with poly2 forming the top plate. The chip has 40 pins and occupies 2.1 x 2.1 mm^2 of silicon area. The prototype network has three inputs while the output layer has 10 neurons. The neurons and synaptic connections are shown in Fig. 6. Therefore, the fabricated chip contains 10 neurons of Fig. 2(b) and 30 synapses. Each synapse is made up of Figs. 3(b) and 4.

Fig. 7 shows the photograph of the fabricated chip and the chip I/O is summarized in Fig. 8. All signals, except four as indicated in Fig. 8, are digital. Since there are 30 synapses, three ROW and 10 COLUMN inputs are provided in order to access the storage capacitor of individual synapse. As an example, to access the synapse (either to write the weight or to read the weight for monitoring purpose) that connects input IN1 to neuron OUT2, ROW1 and COLUMN2 should be asserted. Similarly, ROW2 and COLUMN8 must be asserted to access the synapse that connects input IN2 to neuron OUT8.

B. Test Results

To test the PCNN chip, both analog and digital signals are required. All analog inputs, except VREFRESH, are fixed values however. The staircase signal VREFRESH is generated by a DAC. In order to control this DAC, plus generating all the necessary signals, an Altera FPGA board is used. This board in turn is interfaced to a PC through the parallel port [8]. The weight can thus be programmed through a PC. The monitoring of output and weight retention of each synapse can also be done with a lot of flexibility.

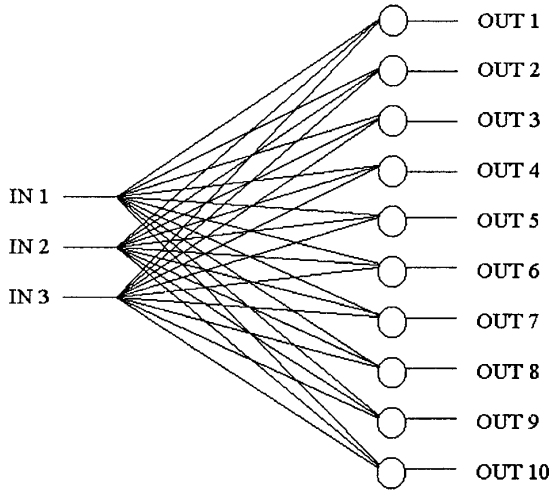

Fig. 6. Prototype neural network.

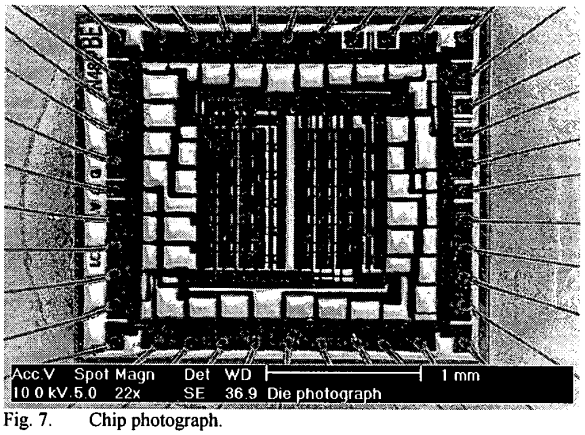

Fig. 7. Chip photograph.

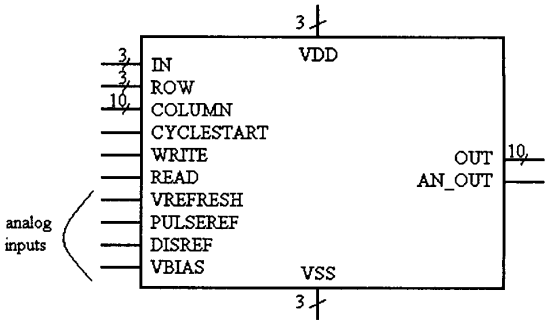

Fig. 8. Chip pinout.

At the moment, a comprehensive test has not been concluded yet. From the preliminary test so far, not all synapses can retain the weight programmed for them. For some synapses, the weight they eventually retain is higher than what they are initially programmed. The exact nature of this defect, and their causes, is still under investigation.

A good result is obtained regarding the effect of VBIAS on the pulse duration of monostable output. This output is manifest in terms of chip output AN_OUT (see Fig. 4.) As can be seen from Fig. 5, when M9 is turned on by the monostable output, VBIAS input controls the rate of parasitic capacitor discharge at the Schmitt trigger input. The higher VBIAS is, the faster the discharge process and hence the quicker for the Schmitt trigger input to reach the lower threshold voltage. In short, a higher VBIAS reduces the pulse width, and vice versa. The relationship is confirmed by the test data, which is shown in Fig. 9.

V. CONCLUSIONS

A PCNN has been fabricated and tested. It consists of a 3 x 10 array of neurons connected by 30 synapses. The design of major components in the neural network chip has been elaborated. The weight for each synapse is stored on a capacitor. Due to inherent current leak in MOS technology, a rather complicated refresh mechanism to retain the correct weight is employed. This external circuit has been constructed on an FPGA board and controlled through PC parallel port. From preliminary tests,

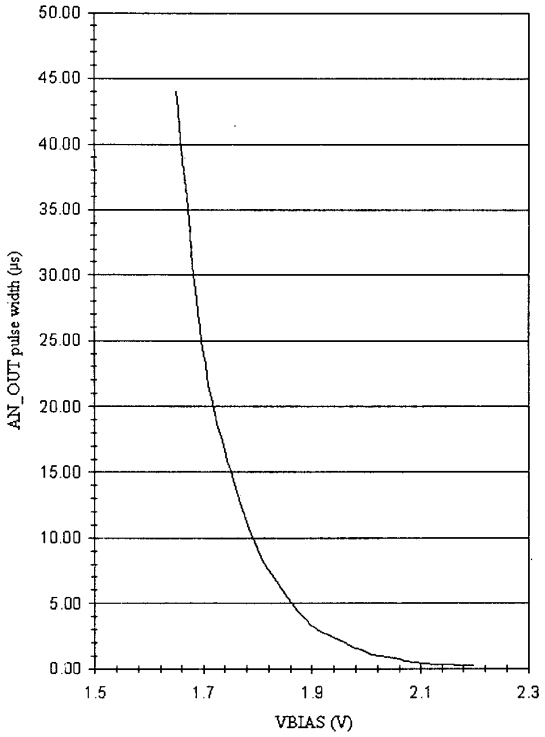

Fig. 9. Control of monostable output pulse width by VBIAS.

there appears to be some problem with synapse weight retention. A more thorough testing to find exact nature and causes of the problem is still underway. A Winner-Take-All PCNN is more useful in terms of solving classification problems and thus should be implemented. Only a minor modification to the neuron is needed for this WTA implementation concept.

ACKNOWLEDGEMENT

The authors would like to thank Astuty Amrin and Nazri M. Ali from Materials Science Laboratory, Fakulti Kejuruteraan Mekanikal, Universiti Teknologi Malaysia for taking the SEM photograph of the fabricated chip.

REFERENCES

[1] A. F. Murray, D. Del Corso, and L. Tarassenko, "Pulse-Stream VLSI Neural Networks Mixing Analog and Digital Techniques," *IEEE Trans. Neural Networks*, vol. 2, pp. 193-204, Jan. 1991.

[2] J. L. Meador *et. al.*, "Programmable Impulse Neural Circuits," *IEEE Trans. Neural Networks*, vol. 2, pp. 101-109, Jan. 1991.

[3] B. Fowler and A. El Gamal, "Pulse-Modulated Analog Neuron Circuits," *Analog Integrated Circuits and Signal Processing*, vol. 6, pp. 45-51, 1994.

[4] A. Mortara and E. Vittoz, "A Communication Architecture Tailored for Analog VLSI Artificial Neural Networks: Intrinsic Performance and Limitations," *IEEE Trans. Neural Networks*, vol. 5, no. 3, pp. 459-466, May 1994.

[5] D. Watola and J. L. Meador, "Competitive Learning in Asynchronous Pulse Density Integrated Circuits," *Analog Integrated Circuits and Signal Processing*, V.2, pp. 61-82, 1992.

[6] P. D. Hylander, "Design Automation Tools for Pulse Coded VLSI Neural Networks," M.Sc. thesis, Washington State University, Pullman, WA, 1994.

[7] P. Allen and D. Holberg, *CMOS Analog Circuit Design*. Fort Worth, TX: Holt, Rinehart and Winston, 1987.

[8] C.-W. Po, "Designing of Pulse Coded Neural Network Chip Controller Using FPGA," B.E.E. thesis, Universiti Teknologi Malaysia, Skudai, Johor, 1999.