CMOS Design of Two Winner-Take-All Circuits Using Pulse Duty Cycle Synaptic Weighting

G. Moon
Department of Electronics
College of Natural Sciences
Hallym University
Chunchon 200
Korea (R.O.K.)
(+82) 361-58-1000

M.E. Zaghloul
Electrical Engineering and
Computer Science Department
George Washington University
Washington, DC 20052 USA
(202) 994-3772
zaghloul@seas.gwu.edu

R.W. Newcomb
Microsystems Laboratory
Electrical Engineering Department
University of Maryland
College Park, MD 20742 USA
(301) 454-6869

ABSTRACT

In this paper the pulse coded Neural Processing Element (**NPE**) is used to realize two Winner-Take-All (**WTA**) systems which use an Hebbian type learning rule for setting the synaptic weights. The two systems, temporal and maximum-level, use the same circuit with a small connection change. In the NPE the average pulse duty cycle modulation technique is used to achieve pulse coded weighting for an artificial neuron. The average pulse duty cycle serves as an information mechanism to determine the weight multiplication. SPICE simulations check the theory with a CMOS prototype chip designed and fabricated through MOSIS. Measurements on the chip compared with simulation results verify the operation of the WTA circuitry.

NEURAL PROCESSING ELEMENT WITH PULSE DUTY CYCLE SYNAPTIC WEIGHTING

In [1], [2], [3] an electronic Neural Processing Element (NPE) was developed. The synaptic weight in the NPE was achieved using Pulse Duty Cycle Modulation (PDCM). The PDCM allows us to achieve implementation in small size CMOS circuits. The NPE of [1], [2], [3] is composed of four functional blocks: a set of Modified Neural Type Cells (**MNTC**) [one for each input], summation, threshold logic, and optional learning blocks [one for each weight]. Figure 1 shows the functional block diagram of the NPE.

Each input signal is multiplied by a corresponding weight in its MNTC. An MNTC functions as a synaptic junction, and accepts analog continuous signal X(t) as the input and analog continuous signal W(t) as the weight. The output of the MNTC is a pulse stream Y_p with a pulse duty cycle that is monotonically proportional to the input signal X(t), and to the input weight W(t). For the output pulse stream Y_p, the Pulse Duty Cycle (**PDC**) is defined as

$$PDC = \frac{\sum_{j=1}^{m} PW(j)}{t} \qquad (1)$$

where PW(j) is the jth pulse width in the stream, assumed to be of m pulses in time interval t. Several pulse streams are summed in the summation block (shown in Fig. 2) using charge accumulation in a conventional capacitor C_S, as also shown in Fig. 2.

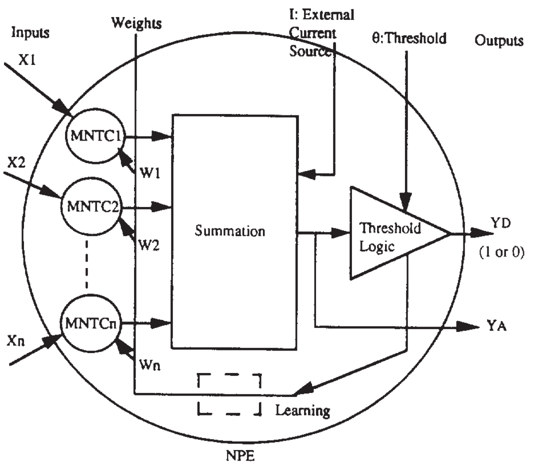


Figure 1. Functional block diagram of Neural Processing Element (NPE).

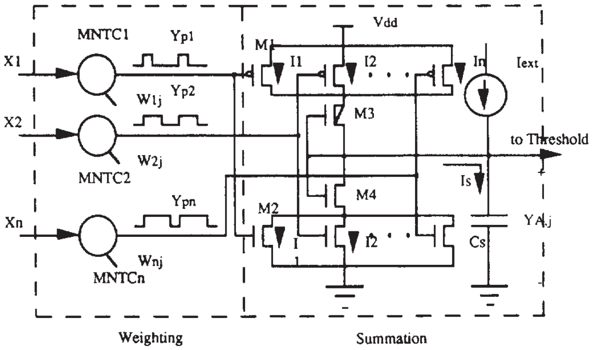


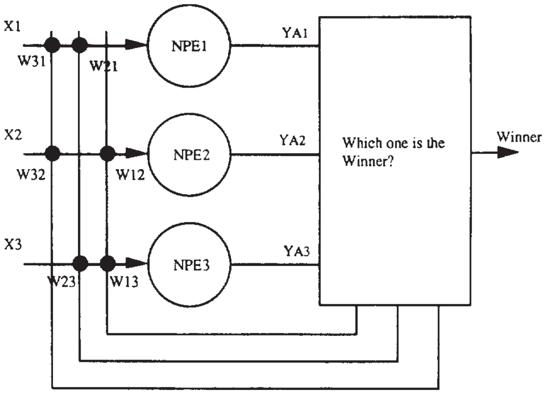
Figure 2. Summation block with n-MNTCs.

The output of the summation block, that is, the voltage on the summing capacitor, is an analog voltage Y_A with its value proportional to the average pulse duty cycle of the output pulse streams from the MNTC's. The analog voltage is compared in a threshold logic block (not shown in Fig. 2, but see Fig. 4) with a threshold voltage to produce the digital output Y_D of the NPE.

WINNER-TAKE-ALL MODEL

The WTA is a special case of the continuous Hopfield neural network in which all connections between different neurons are working in the inhibitory mode. Thus, as one of the outputs reaches above the threshold level and fires, it will, as a winner, suppress all the other nodes which then hold their outputs below the threshold level. Depending on how to choose a winner, two types of WTA models are introduced in this section; one being a temporal WTA, where the neuron with an earliest-coming input is to be chosen as a winner in the network, while the second is a maximum-level WTA, where the neuron with the strongest input acts as a winner.

In either case an example of the WTA with three contending inputs is shown in Fig. 3.



● :Synaptic Weights

Figure 3. Winner-Take-All Model with three contending inputs.

A type of Hebbian learning is undertaken [4] for determination of the WTA synaptic weights is taken to be as follows

$$W_{ij} = \overline{Y_{Di}}Y_{Aj}, \text{ for temporal WTA} \qquad (2)$$

$$W_{ij} = \overline{Y_{Di}}X_j \text{ for Max-level WTA} \qquad (3)$$

In (2) and (3) the overbar denotes digital complement (note that the Y_Ds are digital outputs of the NPEs).

As shown in Fig. 2, the net charge across the summation capacitor CS is proportional to the time period when P-type transistor M1 (in Fig. 2) is ON minus the time period the N-type transistor M2 (in Fig. 2) is ON. Both N- and P-type transistors are designed in such a way that when they are turned ON they are operating in their saturation region acting like current sources, and the saturation currents for both P- and N-type transistors are the same. As a result, the summation voltage YAj of each summation block (corresponding to each NPE) is a function of the average PDC of the input streams as defined in equation (1).

The analog voltage YAj can be expressed as [3] as:

$$Y_{Aj} = \frac{Q_o}{C_{Sj}} \sum_{K=1}^{n} \left(1 - 2PDC_K\right) + Y_{Ao} \quad j = 1, \ldots, n \qquad (4)$$

for n inputs to each NPEj, Qo, and YAo are constants. PDCK is the average Pulse Duty Cycle associated with $MNTC_K$'s output streams YPK.

Temporal WTA

For the temporal case, we use eq. (2) as the weight learning rule. Suppose, for example, the first output hits the threshold level making YD1 high. Then according to the learning rule $\overline{Y_{D1}} = 0$ and, the correlate weights, both W12 and W13, become zero inhibiting the other two outputs YD2 and YD3. A single transistor is used for implementing the learning rule. The circuit diagram for the temporal WTA model with three contending inputs is shown in Fig. 4.

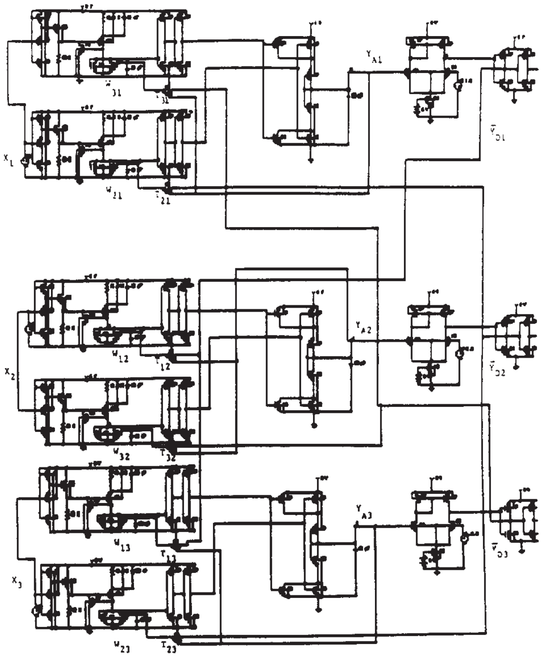


Figure 4. Circuit diagram for temporal WTA.

Three NPEs are used for three contending inputs. Each NPE has two MNTCs for its synaptic junctions connected from the other two NPEs. Notice the learning rule (2) above is accomplished by a single transistor Tij along with a negated digitized output from one NPE and an analog output from the other NPE. The SPICE simulation result is shown in Fig. 5.

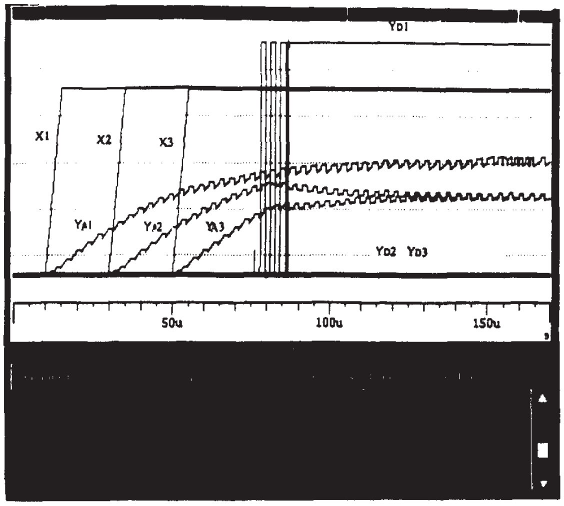


Figure 5. SPICE simulation result for temporal WTA of Fig. 4.

The same level of 4 volts of input are applied to the three NPEs with different delays. The first input comes first delayed only 10us, while the second and the third inputs are delayed 30us and 50us, respectively. As expected and shown, as the output from the first NPE (YA1) reaches the threshold level first, arbitrarily chosen as 2.2 volt, it inhibits the other two outputs (YA2 and YA3) through the changes of their weights as described in eq. (2). As a result, the increases of YA2 and YA3 are significantly degraded and both become saturated around 1.7 volt, while YA1 increases further saturating around 2.5 volt. The winner of this model is therefore the first NPE which has the earliest-coming input and the simulation shows the correct result.

Max.-Level WTA

For the Max-level WTA case, the learning rule is expressed by (3). The first term $\overline{Y_{Di}}$ is for suppressing the other weights in case the i-th output is HIGH as a winner. The second term X_j reflects the effect of the input strength. Thus, the weight between i-th and j-th NPE (Wij) is proportional to the negated digitized output of one NPE as well as the input of the other. Thus, if one neuron with lower (weaker) level of input claims that it is the winner just because it has earliest-coming input, then this input term (Xj) in the learning rule will play a role to boost the weights of the neuron with the maximum strength, and eventually let the network choose it as a winner. For example, in the two neuron case, if both YD1 and YD2 are low at the beginning, and if X1 > X2 the weight W21 is larger than W12 and this will boost the output of the first neuron, which is to be a winner. The circuit diagram for this case is similar to that in Fig. 4. However, the connection for the learning rule of (3) will be different from Fig. 4, where each transistor Tij is connected to the input Xj and the negated digitized output $\overline{Y_{Di}}$. Simulation results for this case are shown in Fig. 6.

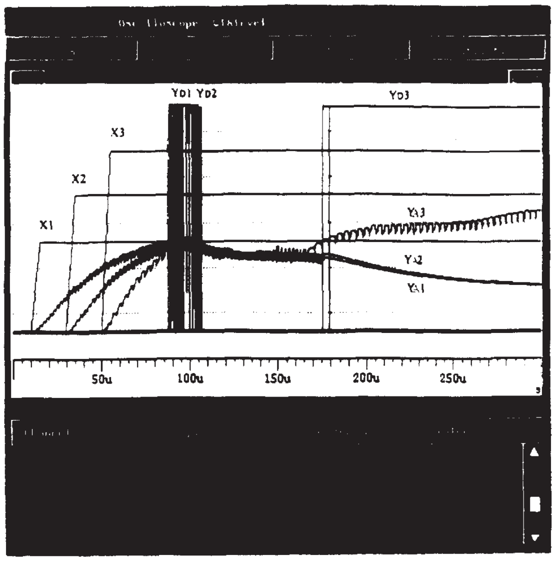


Figure 6. SPICE simulation result for Max-level WTA.

Three inputs X1, X2, and X3, of 2V, 3V, and 4V, respectively, are applied to three NPEs with different delays. X1 input comes first, while X2 is applied 20usec behind the X1, and X3 is applied 20usec behind the X2. The threshold θ is fixed as 2.2V for all NPEs. This threshold value is chosen arbitrarily near half of Vdd, and should not be too close to GND level. If the threshold level is too small, the network decides the winner too soon before it receives all inputs and thus chooses a wrong winner. If the threshold is large, the model works correctly, since the input with maximum level reaches the threshold level first. So, the threshold should be chosen preferably near half of Vdd or higher. Figure 6 is for a worst case scenario when the threshold (2.2V) is below half of Vdd (5V). Although X3 is applied with the longest delay the system is to select the third output (YD3) as a winner because X3 has the maximum strength (level). Around 100usec, the network falsely chooses the first output (YD1) as a winner because it reaches the 2.2V threshold level. Soon after the second output (YD2) is also chosen as a winner for a short period of time, and thereafter until at 170usec the three neurons 'fight' each other to claim a winner. The output for the third neuron, however, is strongly increasing over the other outputs due to the learning rule of eq. (3), and as a result, at 180usec and, thereafter, the network finally chooses YD3 as the winner, which is the correct answer.

VLSI CHIP DESIGN AND MEASUREMENTS

With circuit diagrams along with design parameters proven through simulation, a CMOS chip was built for hardware implementation. With this chip, we are able to test and measure the basic functions of the NPE as well as the WTAs described in the previous section. The chip design and fabrication was done through MOSIS. CMOS double metal P-well technology was adopted for the fabrication. The minimum feature size is 2um. Process parameters as well as some electrical parameters had been given by MOSIS which were used for the simulations in the previous sections. The TinyChip has 40 pins and is packaged in a ceramic DIP (Dual In-line Package). From the MOSIS original design, six pins are pre-assigned (hardwired) for power connections: three for Vdd, and three for GND, and thus only the remaining 34 pins are available for user connections. Due to this pin count restriction, a modification was made on the original frame so that 38 usable pins became available along with 2 power pins. In the following section measurements of the temporal WTA circuit are illustrated to verify the above design.

Measurements of Temporal Winner-Take-All

For the Temporal WTA, we need to set up an adaptive learning feedback which will execute the learning rule of eq. (2), $W_{ij} = Y_{Di}Y_{Aj}$. This was done using single 50/50um N-type transistors in the learning block as explained in the previous section. For Temporal WTA, we also used two off-chip TTL 7404 (inverters) packages for introducing the delay between two inputs. Six inverters in two 7404 TTL packages were connected in cascade for this delay.

Two unit-step signals with a delay between them were applied as two inputs, and two digitized outputs (YD1 and YD2) were measured. The output associated with the input which comes first is expected to be in the HIGH state and the other output is expected to be LOW. Figure 7 shows the measurement result. Four signals of X1, X2, YD1, and YD2 are shown from the top, respectively. As can be seen, with the six 7404 inverters, 60ns delay was introduced for X2, and as we expected, YD1 was chosen as the Winner because the first NPE has an input which arrives first. Notice YD1 is HIGH while YD2 stays relatively LOW. Notice we have 1V amplitude degradation for X2 since X2 was generated using X1 through six inverters in 7404. Notice also that YD2 is not as close to GND as we would wish. This could be due to many reasons: malfunctioning of the threshold block of the second NPE, for example, or power lines problem in conjunction with TTL chips.

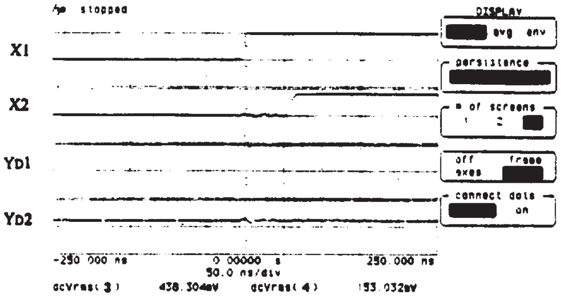


Figure 7. Functional test for Temporal Winner-Take-All model (5V/Vert.).

CONCLUSIONS

Analog CMOS designs of two types of Winner-Take-All circuits are introduced. The WTA circuits use an Hebbian type learning rule for setting the synaptic weights. The learning subcircuit is implemented using a single nMOS device. A CMOS chip was designed and fabricated through MOSIS. Measurements check the proposed design. In addition, as it turns out from the actual layouts, a single MNTC occupies 150um x 320um (73mil^2) silicon area. Thus, we will be able to build up to 30 MNTCs on MOSIS TinyChip 2.1mm x 2.1mm, which can function as five fully-connected neurons. This is an expected value based on an assumption that 30% of the chip will be used for routing (compared with 50% in single-poly single-metal technology), and another 20% reserved for the other blocks, like summations and thresholds. Thus the design introduced occupies a reasonably small silicon area.

REFERENCES

[1] G. Moon, M.E. Zaghloul, and R. Newcomb, "An Improved Neural Processing Element Using Pulse-Coded Weights," *Proc. IEEE ISCAS*, Chicago, pp. 2760-2763, May 1993.

[2] G. Moon, M.E. Zaghloul, and R.W. Newcomb, "A CMOS Implementation of Neural Processing Element with Pulse Duty Cycle Synaptic Weighting," in preparation.

[3] G. Moon, "VLSI Design of Neural Networks Using Pulse Coded Weights With On-Chip Learning Capability", Ph.D. Thesis, Department of Electrical Engineering and Computer Science, The George Washington University, Washington, DC, May 1993.

[4] D. Hebb, The Organization of Behavior, John Wiley & Sons, New York, 1949.