A 1 MHz Compact Digitally Controlled Perceptron Integrated Circuit Implementation with Process Insensitivity

José L. Pino, Terry L. Sculley and Martin A. Brooke

Georgia Institute of Technology, School of Electrical Engineering, Atlanta, Georgia 30332
Telephone: (404) 894-3304

Abstract — A fast, compact process and temperature insensitive perceptron neural network is introduced. The integrated circuit proposed forms one layer in a perceptron neural network that can be easily cascaded to form a multiple layer network. The implementation realizes practical eight bit two's complement digital weights.

Perceptron neural networks are capable of various pattern and image recognition tasks. Single layer perceptron neural networks can be easily taught to recognize simple patterns, while their multiple layer counterparts are capable of recognizing much more complicated patterns using more sophisticated training algorithms [1,2]. This paper introduces a single layer perceptron neural network integrated circuit capable of being cascaded in series to form a multiple layer network. The implementation's advantages over previous designs are its compactness, speed, unique weight structure, and temperature and process insensitivity.

Each layer of the perceptron network has fully interconnected weighted inputs feeding into activation amplifiers. The network relies on current rather than voltage signals, eliminating

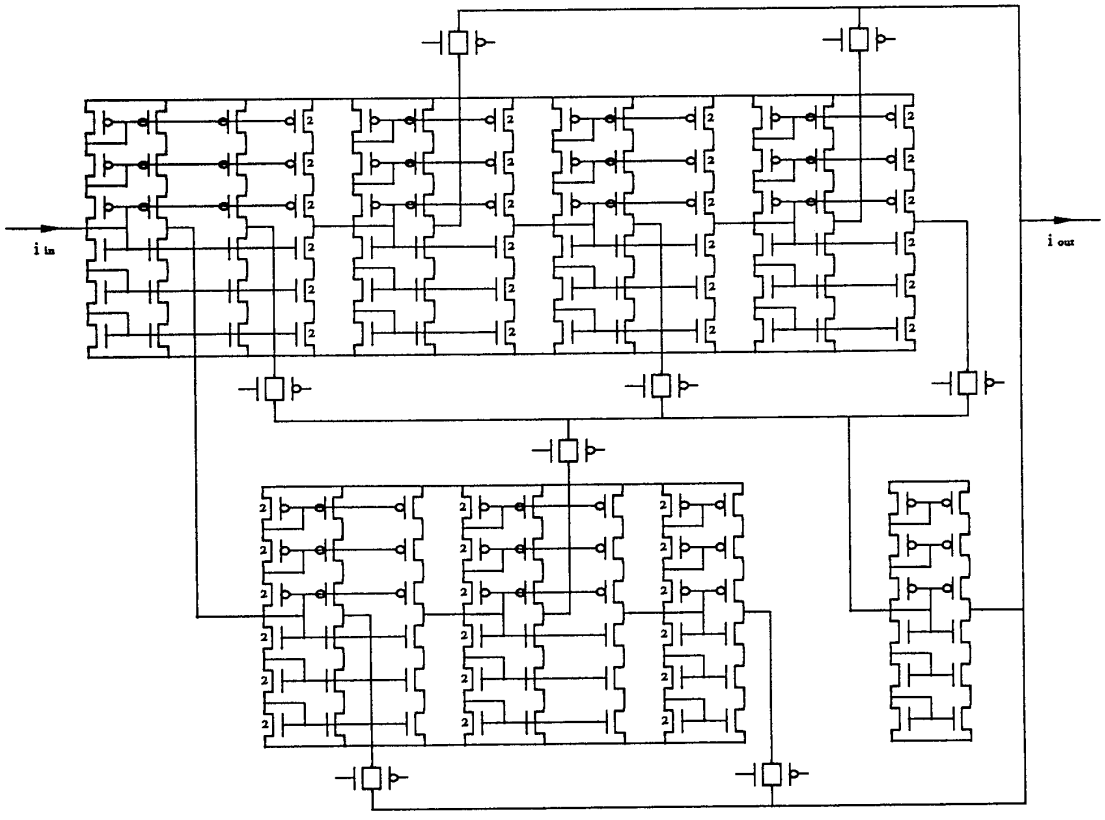

Figure 1. A single eight bit digitally adjustable current mirror weight stage

the high impedance nodes seen in many other perceptron implementations. Since this network uses current signals it is possible to achieve a design using mostly current mirrors. Current mirrors are both highly temperature and process insensitive. The overall gain of the circuit in the active region is unity. The vast majority of the CMOS components utilized are of minimal size (3μ/2μ) with a few exceptions seen in the weight stage and the reference current mirror stage.

Once a current is fed as an input, it is amplified by a digitally programmable current mirror. The amplification is accomplished through cascaded current mirrors, allowing eight bits of resolution while still maintaining a relatively compact circuit (see figure 1). Past digital weight implementations have been cumbersome because of the use of parallel rather than series weight circuits. Since eight bit programmable weights are sufficient for many neural networks applications, a eight bit two's complement coding structure is implemented [3]. Weight accuracies in excess of eight bits are limited by device mismatch. The weights are digitally adjustable in multiples of 0.125 over a range from -16.000 to 15.875.

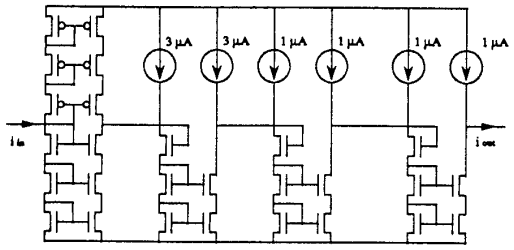

Figure 2. Activation Amplifier

The weighted inputs are then summed to form a net excitation. Since current signals are used, this is easily accomplished by feeding all the excitations into a common node. The net excitation is then fed into a activation amplifier which clips the signal to within an adjustable range. The activation amplifier is composed of four stages: the first stage functioning as a buffer, the second stage as a level shifter, and the last two stages as limiters, clipping the signal to

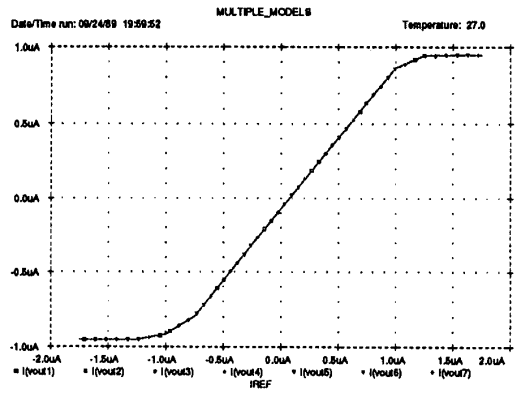

Figure 3. Sigmoid transfer characteristic of the activation amplifier

within a specified range (see figure 2). The ideal current sources shown are actually a mirrored reference current. The output wave form can vary from a nicely rounded sigmoid (see figure 3) to a piecewise linear function (see figure 4) depending on the clipping range specified. The rounding seen in figure 3 is due to subthreshold currents. In figure 4, these subthreshold currents become insignificant relative to the output signal.

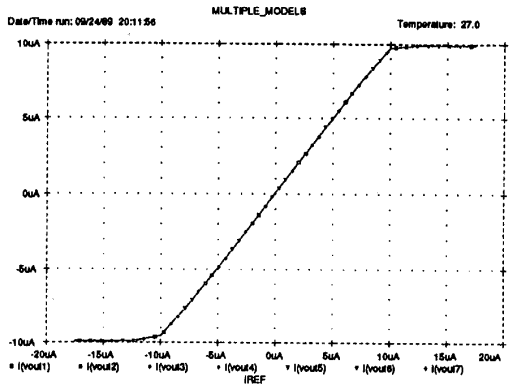

Figure 4. Piecewise linear transfer characteristic of the activation amplifier

Both figure 3 and figure 4 are plots from seven distinct MOSIS model parameter runs superimposed on top of each other with some parameters varying up to an order in magnitude [4]. As can be seen from these plots, this nonlinearity circuit presented is process insensitive, making the output wave forms from

different integrated circuits uniform. Many other implementations utilize a CMOS inverter to obtain a sigmoid response, making the output wave form process sensitive, i.e. making it unpredictable from one integrated circuit to another [5].

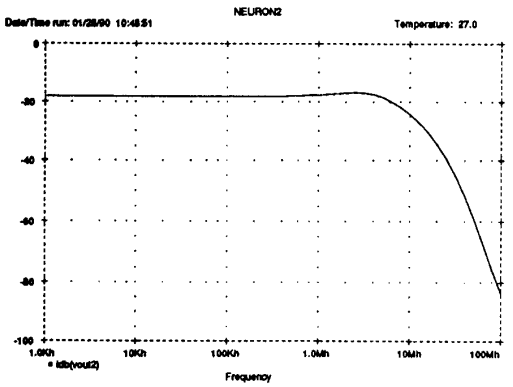

Figure 5. Transfer function of a weight circuit in series with an activation amplifier

Based on simulation runs the design presented is capable of operating at speeds up to 1 MHz (see figure 5). Given this speed and the high predictability of the implementation, numerous applications are possible. One such application is a real time image compressor/ expander. On the front end a two stage perceptron network compresses the analog image data while at the opposite end another two stage network decodes the data. Since the wave forms are predictable, the image data would transmit virtually undistorted. Applications using the design presented could be mass produced and programmed in house with reasonable certainty of their individual transfer characteristics.

The perceptron implementation presented holds promise in that it has a highly predictable output independent of process parameters and thermal effects. Other strengths in this design are its speed and the fact that each weight is fully programmable with up to eight bits of resolution while maintaining a compact size. Each integrated circuit will comprise a single interconnected layer capable of being cascaded to form a multiple layer network. The proposed neural network integrated circuit could be easily mass produced and potentially has a wide range of applications, such as video signal compression/expansion.

References

[1] R.Rosenblatt, Principles of Neurodynamics. New York: Spartan Books, 1959.

[2] D.E.Rumelhart, G.E.Hinton, and R.J.Williams, "Learning Internal Representations by Error Propagation," Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations. MIT Press, 1986.

[3] S.L.Gilbert, "Implementing Circuitry: A Design Proposal for Back-Propagation," Master's Thesis MIT, May, 1988.

[4] MOSIS, Marina Del Rey, Los Angeles, California, Parameters available from MOSIS - Run Names: M73U, M77P, M79X, M7C0, M84N, M85T, M86Y.

[5] H.P.Graf and R.E. Howard, "Electronic Neural Network Chips," Applied Optics, Vol. 26 No. 23, December 1, 1987.