An Electrically Trainable Artificial Neural Network (ETANN) with 10240 "Floating Gate" Synapses

Mark Holler, Simon Tam, Hernan Castro, Ronald Benson*
Intel Corporation, Technology Development
Novel Device Group
2250 Mission College Blvd. SC9-34
Santa Clara, Ca. 95052-8125
Ph. 916-351-2704

ABSTRACT

Use of "floating gate" non-volatile memory technology for analog storage of connection strengths or "weights" has previously been proposed and demonstrated. This paper reports the analog storage and multiply characteristics of a new floating gate synapse and further discusses the architecture of a neural network which uses this synapse cell. In the architecture described 8192 synapses are used to fully interconnect 64 neurons and to connect the 64 neurons to each of 64 inputs. Each synapse in the network multiplies a signed analog voltage by a stored weight and generates a differential current proportional to the product. Differential currents are summed on a pair of bit-lines, transferred through a sigmoid function and appear at the neuron output as an analog voltage. Input and output levels are compatible for ease in cascade connecting these devices into multi-layer networks. Processing is done in parallel with an anticipated delay from input to output of approximately 1us. Weights are changed individually similar to the technique used to write data into EEPROM's. The pulse width and height of a weight change pulse must be calculated externally. Synapse cell size is 2009 sq. microns using a 1u CMOS EEPROM technology.

INTRODUCTION

Implementation of modifiable artificial neural connections or synapses using "Floating gate" non-volatile memory technology [1], has been proposed by Alspector, et al [2]. Alspector proposed the use of an analog charge on a floating gate to replace digital flip-flops used to store a connection strength or weight. The purpose of substitution being to reduce the size of the synapse. Since this time several specific floating gate synapse circuits have been revealed [3,4,5]. Electrical results have been reported for one of these cells [3]. This paper describes a new floating gate synapse cell which operates differentially to avoid the problems associated with power supply and temperature changes which occur between the time a network is trained and the time it processes information. Differential operation affords several other benefits such as full four quadrant multiply function and improved linearity of the multiply.

The cell described places a variety of constraints on the architecture of the network in which it is used. The architecture of an Electrically Trainable Analog Neural Network (ETANN) which uses the synapse described will be discussed to elaborate on some of the issues encountered in using a fully differential, analog, non-volatile synapse. General issues pertaining to VLSI implementation of neural networks are also discussed. Some key aspects of the architecture are: emphasis on speed via parallelism, both feed forward and feedback connections on chip, analog voltage input and output, a large pin count package, and the option to operate in a fast static mode or in a clocked mode useful with multiplexed external buses.

Learning calculations have been left off chip to allow flexibility in applying the network and to avoid the complexity and cost penalties associated with implementing learning on chip. Writing weights into the ETANN will be referred to as training rather than learning as it implies passive involvement of the ETANN. Several additions have been made to the architecture to speed up training and minimize the connection overhead required by external learning and control of training. Addresses are provided which allow the external processor to select individual weights for modification in the same way that a byte of data is addressed in an EEPROM.

SYNAPSE CELL

The synapse cell circuit, shown in Figure 1, is an NMOS version of a Gilbert-Multiplier [6]. A pair of EEPROM cells are incorporated in which a differential voltage representing the weight may be stored or adjusted.

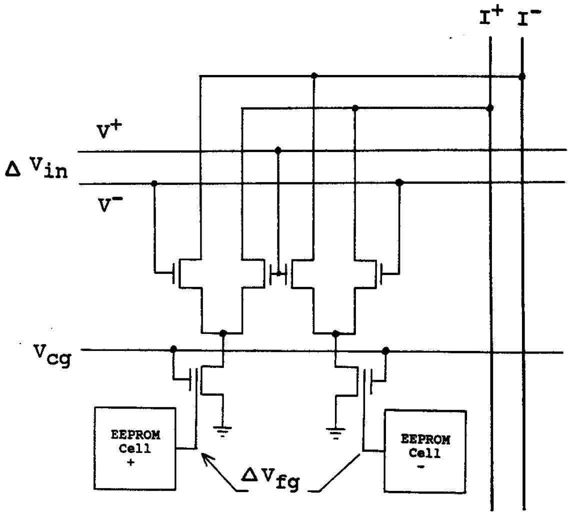

Figure 1. Differential "floating gate" synapse schematic diagram.

Electrons are added to or removed from the floating gates in the EEPROM cells by Fowler-Nordheim tunneling of electrons between the floating gates and the diffusions. A desired differential floating gate voltage can be attained by monitoring the conductances of the respective EEPROM MOSFETs. In particular, the respective floating gate voltages may be expressed as:

$$V_{fg} = \frac{C_{pp}}{C_{tot}} * V_{cg} + \frac{Q_{fg}}{C_{tot}}$$

where C_{pp} is the capacitance between the floating gate and the top gate (control gate) and C_{tot} is the total floating gate capacitance. V_{cg} is a static control gate voltage used for biasing and Q_{fg} is the charge stored on the floating gate. Assuming that all the transistors in the MOS multiplier stayed in saturation, we have the differential output current as:

$$\Delta I_{out} = I^{+} - I^{-}$$

$$\Delta V_{fg} = \frac{\Delta Q_{fg}}{C_{tot}}$$

$$\Delta I_{out} = \Delta V_{in} * \Delta V_{fg}$$

where ΔV_{in} is the differential input voltage and ΔQ_{fg} is the differential charge stored between the two EEPROMs' floating gates.

A photo-micrograph of the synapse circuit is shown in Figure 2. Figure 3 shows the output characteristics of the synapse with symmetrical differential weights. The differential input voltage plotted along the x-axis is defined to be:

$$\Delta V_{in} = V^{+} - V^{-}$$

In this particular illustration V^{+} was varied from 1V to 5V while V^{-} was fixed at 3V. The differential weight was varied symmetrically between approximately -.8V and +.8V. Since the average of V^{+} and V^{-} increased as V^{+} was varied from 1V to 5V, we observed that the output current ΔI_{out} is higher in the positive ΔV_{in} regime (right half plane) than the negative ΔV_{in} regime.

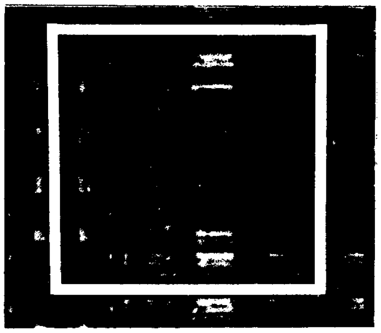

Figure 2. Floating gate synapse cell. Cell size 41.6u x 48.3u.

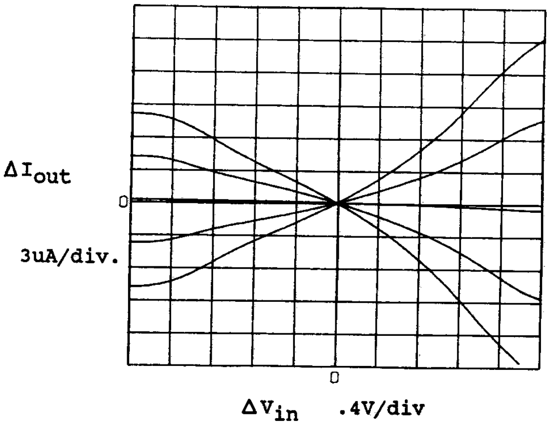

Figure 3. Characteristics of the synapse multiplier for five different values of weight. EEPROM cells written symmetrically.

Figure 4 shows the output characteristics of EEPROM based synapse when the conductance of one of the EEPROM devices was varied through the Fowler-Nordheim tunneling process with the "reference" EEPROM device remaining unchanged. The "weight" EEPROM's threshold was varied from approximately 1.95V below to +1.65V above the "neutral" threshold of the reference EEPROM. These characteristics present how the differential current representing the product changes when the charge on only one floating gate is adjusted.

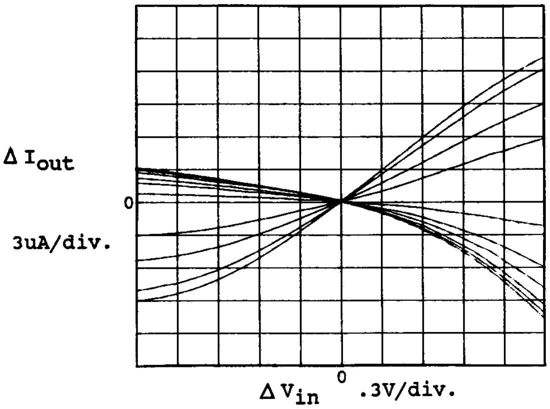

Figure 4. Characteristics of the synapse multiplier for varying values of the weight. One EEPROM cell written.

In previous work [3] it was demonstrated that the voltage on a floating gate could be adjusted with .4% resolution (8 bits) using channel hot-electron programming. The physical limit on the smallest increment which can be made is the voltage shift due to adding a single electron. For the cell discussed in this paper the smallest increment will not be limited by the discreteness of the charge of a single electron but, rather by the ability of the neuron circuits to detect a change. 250 million electrons have to be stored on one of the floating gates to cause a 2V shift in the threshold of the transistor. If single electron increments could be sensed this cell would be capable of storing 30 bits of information.

No data is presented here which shows that single electrons can be added or removed. This is due to measurement limitations not a physical limitation of Fowler-Nordheim tunneling. Fowler-Nordheim tunneling is exponentially dependent upon the voltage applied making it possible to slow the rate of tunneling down indefinitely. Some finite probability of tunneling remains even under normal operating biases. Scaling of floating gate synapse cells will not likely be limited by the discreteness of the electron charge for some time to come.

To show adequate weight change resolution for this cell the threshold of one of the EEPROM devices in a synapse has been plotted as a function of the number of 20us pulses applied. 180 pulses were applied to shift the threshold by 4 V. See Figure 5. The average threshold change per pulse was 22 mV or .6% which corresponds to more than 7 bit weight setting resolution. Higher weight setting resolution is possible.

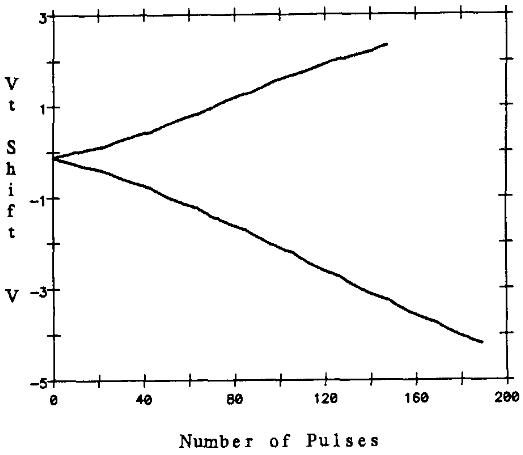

Figure 5. EEPROM threshold vs. the number of 20us pulses applied. Weight change pulse voltage increased by .5 V every 21 pulses.

WEIGHT RETENTION

How well the value of a weight is retained is an important characteristic of any synapse because it limits the amount of information which can be stored in the synapse. For example if a weight changes in time by 10% of full range then no more than 10 different states of the weight can be distinguished from one another. The number of distinct states the weight can have is directly related to its information capacity.

To characterize how much undesirable change in the weights will occur over the life of a device several synapses were baked at 250C to accelerate the charge relaxation phenomenon that is known to occur in floating gate memory devices. Figure 6 shows the percentage change in the weight as a function of

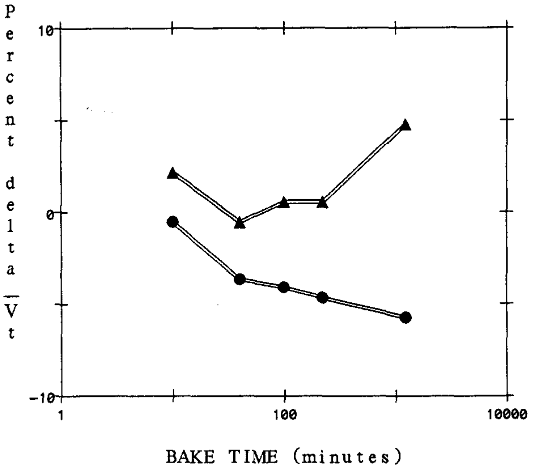

Figure 6. Percent change in EEPROM threshold vs. bake time at 250C. Upper curve - erased cell, Lower curve - programmed cell.

time in the 250C bake. Since no activation energy has yet been determined for the synapse cell it cannot be said with certainty that any point on this graph is a precise equivalent to the desired lifetime of the device. However, using a 1.1eV activation energy derived from a similar floating gate technology [7] a time on the order of 3200 minutes at 250C is equivalent to a 15 year lifetime for a floating gate device at 125C. Extrapolating out to this time in Figure 5 it is indicated that the weight will change approximately 6-7% which would allow long term storage of one of 14-17 distinct states. Sixteen distinct states corresponds to 4 bits of resolution.

In applications where the network is retrained frequently the high weight setting resolution of the cell can be utilized. In applications where the network is trained only once and used for a long period of time 4-bit resolution appears to be the limit. However, it may be possible to accelerate the relaxation process and retrain the network or pre-compensate for the relaxation phenomenon and bake the network to improve resolution.

NETWORK ARCHITECTURE

Learning vs. Speed

Two of the desirable characteristics of neural networks are their ability to learn and their potential for high speed processing when truly parallel architectures are implemented. The architecture which will be discussed is optimized for processing speed. Learning capability is provided but, only in a rudimentary sense.

For maximum processing speed as much parallelism as possible was attempted, multiplexing was avoided. The limiting factor in the amount of parallelism which could be achieved was the pin count of practical integrated circuit packages. A large pin count pin grid array package was selected for this reason.

Analog input and output were selected to

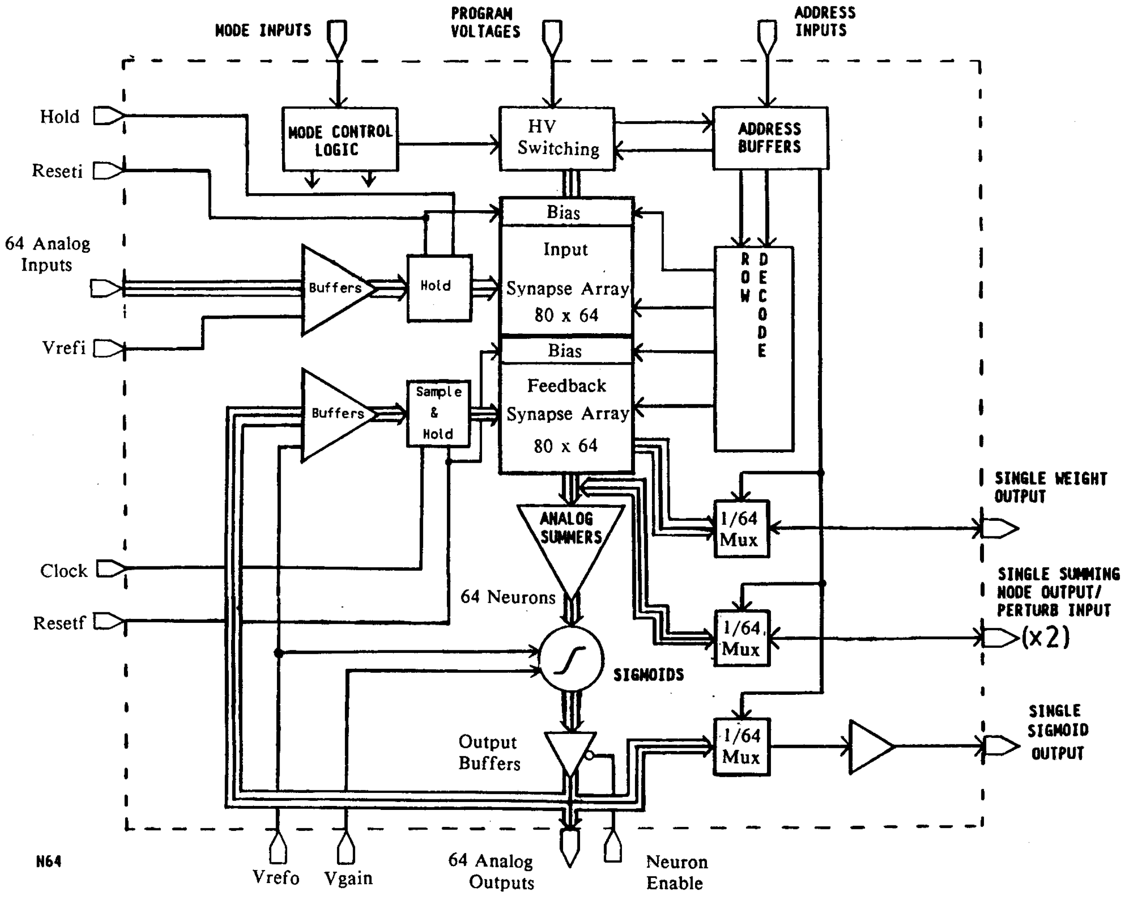

Figure 7. ETANN Block diagram

maximize information throughput on the limited number of pins.

An external intelligent supervisor must calculate the weight changes and must calculate the voltages to be applied to each synapse to modify the weight. Weights are addressed and are changed one at a time. The learning rate will likely not be any faster than an average simulation. However, one benefit of this approach is that the network can be trained equally well by a training controller using Hebbian learning, back-propagation, unsupervised learning, Madaline or nearly any other learning algorithm. It was necessary to add circuitry to be able to perturb the neuron inputs as is necessary in Madaline learning.

Architecture with Feedback

A block diagram of the architecture of the ETANN is shown in Figure 7. The chip has 64 inputs, and 64 neurons fully interconnected by 4096 synapses. Neuron outputs are available at 64 output pins.

A 64 signal feedback path from the outputs to a separate synapse array of 4096 synapses was provided. It allows implemention of Hopfield networks [8] and networks capable of processing and producing sequences of patterns [9]. The die size did not increase signicantly with the addition of the feedback synapse array since the die size is limited primarily by the perimeter required for the large number of bonding pads. See Figure 9.

Feedback is gated by a clock controlling the sample and hold buffers in much the same way that feedback is gated by a clock in synchronous digital circuits. "Free-running" Oscillations are prevented regardless of the values given to the weights in the feedback synapse array. A repeating sequence of patterns can still be generated at the neuron outputs but the rate of repetition is controlled by the clock.

The sample and hold buffer is also useful for storing data input via the output pins. A signal is provided which can disable the neuron output buffers, allowing data to be input to the chip via the output pins. In this mode of operation input pattern vectors with 128 components may be processed.

The ETANN has a single layer of neurons for maximum flexibility in design of multi-layer networks at board level. However, multiplexing the use of the neurons is allowed which makes two layer operation of a single ETANN possible. Two layer operation is discussed below.

Extra rows of synapses are provided with fixed positive input to allow setting a bias for each neuron. A block of 16 rows of biasing synapses are enabled when the input connections are enabled and a separate block of 16 rows is enabled when the feedback connections are enabled. A total of 2048 synapses are dedicated to setting biases.

Two Layer Operation

Two layer operation is accomplished by first disabling the feedback synapse array and applying 64 input signals to the ETANN.

After sufficient time is allowed for the outputs to settle the outputs are clocked into the sample and hold buffer in the feedback path. Next the input synapse array is disabled which allows the stored previous state of the outputs to be processed without interference. The feedback synapse array is used as a second layer of connections. The neuron amplifiers which are a significant portion of the die area are used twice in two layer processing.

Differential circuitry was used throughout the analog signal path for power supply and temperature immunity as well as speed of operation.

Scaled CMOS technology is used for its packing density, performance and it's ability to power down to very low power levels.

INPUT/OUTPUT

Analog input and output was chosen over digital to maximize the information carrying capacity of the interconnect. Single ended inputs rather than differential inputs were also chosen to conserve pins. Each input is converted to a differential signal in the first stage of the input buffer which is a differential amplifier. The other input to the differential amplifier is a reference voltage supplied by the user. This reference voltage can have any value from 0 to 1.7V. Voltages above the reference are positive quantities while voltages below the reference are negative quantities for purposes of the multiply which occurs in the synapse. If the reference voltage is set to zero the inputs can only take on positive values. And the synapse is used as a 2-quadrant multiplier rather than a 4-quadrant multiplier. If the reference voltage is set to 1.5V the inputs become TTL compatible.

The use of a differential synapse to reduce the sensitivity to voltage and temperature variations yielded a synapse with full 4-quadrant multiply capability. Both the inputs and the stored weights can be positive or negative and the synapse will produce the correct signs for the products in all cases. The inputs may take on inhibitory characteristics and if connected to an inhibitory connection (a synapse storing a negative weight) will cause excitation of the subsequent neuron. Although this type of inhibitory input/inhibitory connection is generally thought not to exist in biological systems there is an indication in simulation results that it can speed up learning.

The neuron amplifiers sum the differential currents produced by the synapses and produce a differential voltage which is fed through an electronic emulation of a sigmoid function. The sigmoid function has variable gain controlled by the voltage applied to one external pin. Simulations of the sigmoidal transfer characteristics with various levels of gain are shown in Figure 8. The output inflection point is controlled by a reference voltage and can be matched with the input zero point by connecting the input and output references to the same voltage.

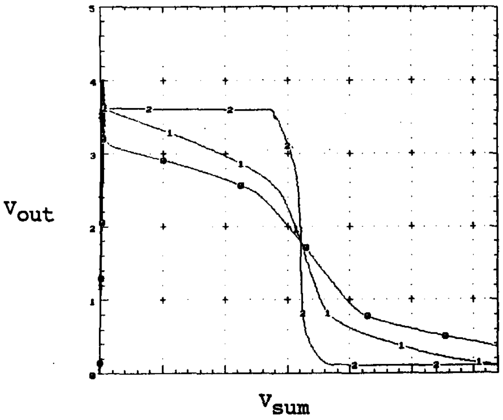

Figure 8. Simulated output characteristics as a function of the total sum of products V_{sum}. Characteristics for three different gain settings shown.

CHANGING WEIGHTS

Modifying weights is done much like programming an EEPROM except that the goal is to end up with a relatively precise amount of charge on the two floating gates in the cell rather than just more than some fixed value as in EEPROMs. To accomplish this an iterative process will likely have to be used. First the present state of the EEPROM cell floating gate is measured by using the multiplexer shown in the lower right hand corner of Figure 7. An address is applied to the 14 address lines to select one of the 10240 synapses. The external processor directing the training calculates the desired weight change based on the network's performance to the training pattern set and then calculates the appropriate pulse height based on the present state of the weight and the desired weight change. The weight can be measured again immediately and if further adjustment is required, another pulse of appropriate height can be applied. Each pulse will be in the range of 10us to 1ms in length and 12-20V in height. Because EEPROM cells are used the weights may be changed in either direction, thus, overshooting a desired value is not a problem.

An output multiplexer was provided to aid in reducing the number of connections the training processor will need to access information at the outputs. The neuron outputs can be read individually at a single output by applying an address to the 6 low order address pins and measuring the output voltage at the pin labeled single sigmoid output in Figure 7.

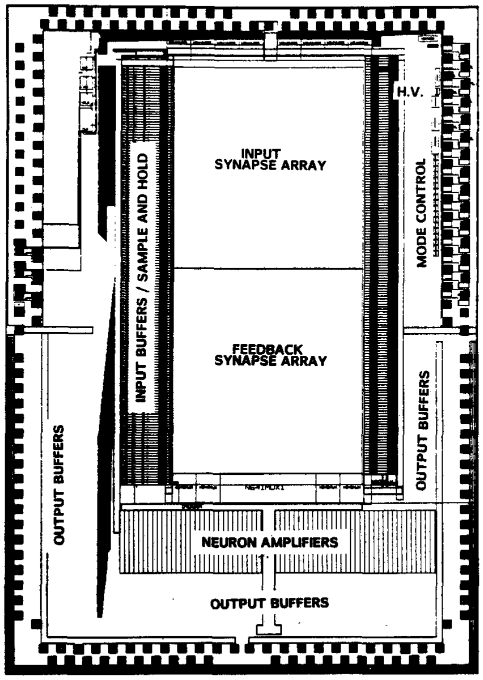

Figure 9. Chip plan of the ETANN discussed.

CONCLUSIONS

Analog weight storage and retention as well as 4-quadrant multiplication in a floating gate synapse cell have been demonstrated. An ETANN which uses 10240 of these cells has been designed and is anticipated to be capable of performing on the order of 10^{10} low precision multiplications per second relevant for pattern recognition tasks. This is several orders of magnitude more processing performance than Von Neumann computer based neural network simulations. Although learning speed had to be compromised significantly to accomplish this performance, this choice was made because it moves computing into a new realm of performance. Had the architecture been optimized for learning speed it is unlikely that the network would learn any faster than fast Von Neumann based computer simulations due to the relative slowness of the Fowler-Nordheim tunneling process and the difficulty of applying many different high voltages to the synapses in parallel. For this reason it seems apparent that floating gate silicon implementations of neural networks provide the most significant new capability by being optimized for speed.

ACKNOWLEDGEMENTS

The authors would like to thank Bruce McCormick and George Canepa for introducing us to neural networks and Dan Seligson for balancing our perspective by championing digital and difference calculating synapse approaches. Special thanks also to FAB 1 and the Flash Memory Group for fabricating the first synapses so quickly. This effort has been assisted by continued technical discussions with the Naval Weapons Center, China Lake in accordance with their need and request for VLSI ANN development or devices as specified in BAA# N60530-88-R-0427. *This material is based partially upon work supported under a National Science Foundation graduate fellowship. Ron Benson is a graduate fellow at the California Institute of Technology.

REFERENCES

[1] Frohman-Bentchkowsky, D., "Memory Behavior in a Floating-Gate Avalanche-Injection MOS (FAMOS) Structure", Applied Physics Letters, Volume 18, Number 8, April 1971.

[2] Alspector, J., et al, "A Neuromorphic VLSI Learning System", Proc. of the 1987 Stanford Conf., Advanced Research in VLSI, pp 313-349, 1987

[3] Tam, S., Holler, M.A., Canepa, G., "Neural Network Synaptic Connections Using Floating Gate Non-Volatile Elements", Neural Networks for Computing, AIP Conference Proceedings, Snowbird, Ut., 1988.

[4] Shoemaker, P., Lagnado, I., Shimabukuro, R., "Artificial Neural Network Implementation with Floating Gate MOS Devices", Hardware Implementation of Neuron Nets and Synapses, A Workshop sponsored by NSF and ONR, January 14,15, 1988. San Diego, Ca.

[5] Faggin, F., Lynch, G., Sukonick, J., "Brain Emulation Circuit with Reduced Confusion", U.S. Patent #4,773,024, Issued September 20, 1988.

[6] Soo, D., Meyer, R., "A Four-Quadrant NMOS Analog Multiplier", IEEE JSSC, Vol. SC-17, No. 6, Dec. 1982

[7] Verma, G., Mielke, N., "Reliability Performance of ETOX Based Flash Memories", Procceeding of the 1988 International Reliability Physics Symposium

[8] Hopfield J.J. & Tank D.W., "Computing with Neural Circuits: A Model", Science vol 233, P625, Aug. 86

[9] Kohonen, T. , Self-organization and Associative Memory, Springer-Verlag, 2nd edition, 1988, New York First edition 1984.