Single Transistor Learning Synapse with Long Term Storage

Paul Hasler, Chris Diorio, Bradley A. Minch, and Carver Mead

California Institute of Technology
Pasadena, CA 91125
(818) 395 - 2812
paul@hobiecat.pcmp.caltech.edu

ABSTRACT

We describe the design, fabrication, characterization, and modeling of an array of single transistor synapses. The single transistor synapses simultaneously perform long term weight storage, compute the product of the input and floating gate value, and update the weight value according to a hebbian or a backpropagation learning rule. The charge on the floating gate is decreased by hot electron injection with high selectiviy for a particular synapse. The charge on the floating gate is increased by electron tunneling, which results in high selectivity between rows, but much lower selectivity between columns along a row. When the steady state source current is used as the representation of the weight value, both the incrementing and decrementing functions are proportional to a power of the source current.

I. Introduction

There are five requirements for a learning synapse [1]. First, the weight should be stored permanently in the absence of learning. Second, the synapse must compute as an output the product of the input signal with the synaptic weight. Third, each synapse should require minimal area, resulting in the maximum array size for a given area. Fourth, each synapse should operate with low power dissipation so that the synaptic array is not power constrained. And finally, the array should be capable of implementing either Hebbian or Backpropagation learning rule for modifying the weight on the floating gate.

We have designed, fabricated, characterized, and modeled an array of single transistor synapses which satisfy these five criteria. The single transistor synapses simultaneously perform long term weight storage, compute the product of the input and floating gate value, and update the weight value according to a hebbian or a backpropagation learning rule.

II. Overview

Each synapse in our synaptic array is a single transistor with its weight stored as a charge on a floating silicon gate. Figure 1 shows the circuit diagram of a 2 x 2 array of synapses. The column 'gate' inputs (V_g) are connected to second level polysilicon which capacitively couples to the floating gate. The inputs are shared along a column. The source (V_s), drain (V_d), and tunneling (V_{tun}) terminals are shared along a row. These terminals are involved with computing the output current and feeding back 'error' signal voltages. The FETs are in a moderately doped (1 x $10^{17} cm^{-3}$) substrate, to achieve a high threshold voltage. The moderately doped substrate is formed in the $2\mu m$ MOSIS process by the pbase implant. Each synapse has an additional tunneling junction for modifying the charge on the floating gate. The tunneling juction is formed with high quality gate oxide separating a well region from the floating gate.

The synapse simultanously computes as an output current a product of weight and input signal, and can increment or decrement the weight as a function of its input and error voltages. The particular learning algorithm depends on the circuitry at the boundaries of the array; in particular the circuitry connected to each of the source, drain, and tunneling lines in a row. With charge Q_{fg} on the floating gate and V_s equal to 0 the subthreshold source current is described by

$$I_{synapse} = I_o e^{\frac{Q_{fg}}{Q_o}} e^{\frac{\delta V_g}{U_T}} \qquad (1)$$

where Q_o is a device dependant parameter, and U_T is the thermal voltage $\frac{kT}{q}$. The coupling coefficient, δ, of the gate input to the transistor surface potential is typically less than 0.1. From (1) We can consider the weight as a current, I, defined by

$$I_{synapse} = \left(I_o e^{\frac{Q_{fg}}{Q_o}} e^{\frac{\delta V_{g0}}{U_T}} \right) e^{\frac{\delta \Delta V_g}{U_T}} = I e^{\frac{\delta \Delta V_g}{U_T}} \qquad (2)$$

0-7803-2570-2/95 $4.00 ©1995 IEEE

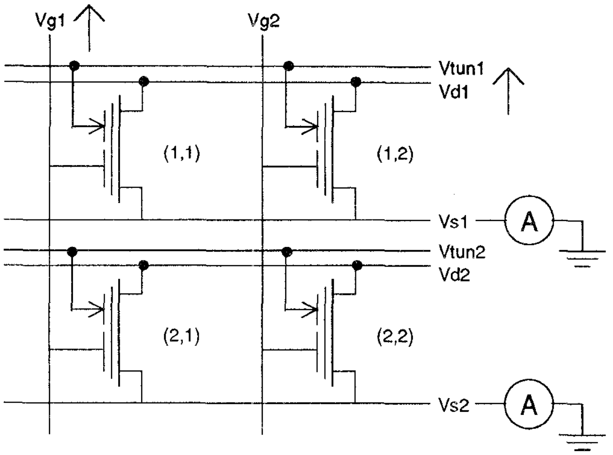

Figure 1: Circuit diagram of the single - transistor synapse array. Each transistor has a floating gate capacitively coupled to an input column line. A tunneling connection (arrow) allows weight increase. Weight decreased is achieved by hot electron injection in the transistor itself. Each synapse is capable of simultaneous feedfoward computations and weight updates. A 2 x 2 section of the array allows us to characterize how modifing a single floating gate (such as synapse (1,1)) effects the neighboring floating gate values. The synapse currents are a measure of the synaptic weights, and are summed along each row by the source (V_s) or drain (V_d) lines into some soma circuit.

where V_{g0} is the input voltage bias, and ΔV_g is $V_g - V_{g0}$. The synaptic current is thus the product of the weight, I, and a weak exponential function of the input voltage. Hot electron injection adds electrons to the floating gate, thereby decreasing the weight. Injection occurs for large drain voltages; therefore the floating gate charge can be reduced during normal feedforward operation by raising the drain voltage. Electron tunneling removes electrons from the floating gate, thereby increasing the weight. The tunneling line controls the tunneling current; thus the floating gate charge can be increased during normal feedforward operation by raising the tunneling line voltage. The tunneling rate is modulated by both the input voltage and the charge on the floating gate.

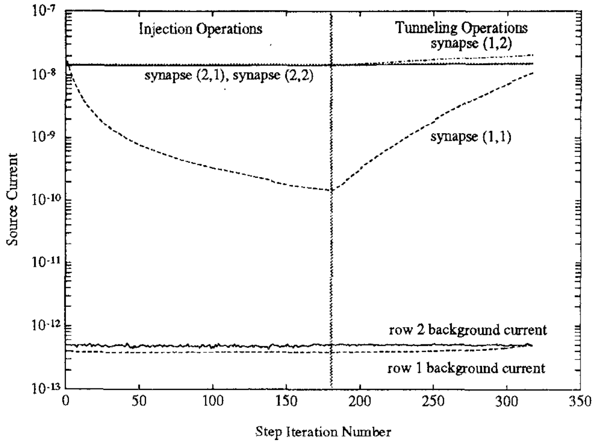

Figure 2: Output currents from a 2 x 2 section of the synapse array, showing 180 injection operations followed by 160 tunneling operations. For the injection operations, the drain ($Vd1$) is pulsed from 2.0 V upto 3.3 V for 0.5s with V_{g1} at $8V$ and V_{g2} at $0V$. For the tunneling operations, the tunneling line ($Vtun1$) is pulsed from 20 V up to 33.5 V with V_{g2} at $0V$ and V_{g1} at $8V$. Because our measurements from the 2 x 2 section come from a larger array, we also display the 'background' current from all other synapses on the row. This background current is several orders of magnitude smaller than the selected synapse current, and therefore negligible.

Figure 2 shows an example the nature of the weight update process. The source current is used as a measure of the synapse weight. The experiment starts with all four synapses set to the same weight current. Then, synapse (1,1) is injected for 180 cycles to preferentialy decrease its weight. Finally, synapse (1,1) is tunneled for 160 cycles to preferentially increase its weight. This experiment shows that a synapse can be incremented by applying a high voltage on tunneling terminals and a low voltage on the input, and can be decremented by applying a high voltage on drain terminals and a high voltage on the input. In the next two sections, we consider the nature of these update functions. In section three we examine the dependance of hot electron injection on the source current of our synapses. In section four we examine the dependance of electron tunneling on the source current of our synapses.

III. Hot Electron Injection

Hot Electron Injection gives us a method to add electrons to the floating gate. First, to inject an electron on the floating gate we need a region where the potential drops more than 3.1 volts in a distance of less than $0.2\mu m$ to allow electrons to gain enough energy to surmount the oxide barrier. Second, we need a field in the oxide in the proper direction to collect electrons after they cross the barrier. The moderate substrate doping level allows us to easily achieve both effects in subthreshold operation. First, the higher substrate doping results in a much higher threshold voltage ($6.1V$), which guarrentees that the field in the oxide at the drain edge of the channel will be in the proper direction for collecting electrons over the useful range of drain voltages. Second, the higher substrate doping results in higher electric fields which yield higher injection efficiencies. The higher injection efficiencies allow the device to have a wide range of drain voltages substantially below the threshold voltage. Figure 3 shows measured data on the change in source current during injection vs. source current for several values of drain voltage.

We can predict the approximate functional form of the curves in Fig. 3 from the device model of the hot electron injector decribed in [3]. Because the source current, I, is related to the floating gate charge, Q_{fg} as shown in (1) and the charge on the floating gate is related to the tunneling or injection current (I_{fg}) by

$$\frac{dQ_{fg}}{dt} = I_{fg} \tag{3}$$

an approximate model for the change of the weight current value is

$$\frac{dI}{dt} = \frac{I}{Q_o} I_{fg} \tag{4}$$

Using a model derived from the Boltzman transport equation [3] the injection current can be approximated over the range of drain voltages shown in Fig. 3 by

$$I_{fg} = -I_s e^{f(V_{d-c})} = -AI_s^{\beta-1} e^{\frac{V_d}{V_{inj}}} \tag{5}$$

where V_{d-c} is the voltage from the drain to the drain edge of the channel, $f()$ is a slowly varying function defined in [3], and V_{inj} is in the range of $60mV$ to $100mV$. A is device dependant parameter. Since hot electron injection adds electrons to the floating gate, the current into the floating gate (I_{fg}) is negative, which results in

$$\frac{dI}{dt} = -A\frac{I^{\beta}}{Q_o} e^{\frac{V_d}{V_{inj}}} \tag{6}$$

The model agrees well with the data in Fig. 3, with β in the range of $1.7 - 1.9$. Injection is very selective along a row with a selectivity coefficient between 10^2 and 10^7 depending upon drain voltage and weight. The injection operations resulted in negligible changes in source current for synapses (2,1) and (2,2).

IV. Electron Tunneling

Electron Tunneling gives us a method for removing electrons from the floating gate. An electric field across the oxide will result in a thinner barrier to the electrons on the floating gate. For a high enough electric field, the electrons can tunnel through the oxide. When travelling through the oxide, some electrons get trapped in the oxide, which changes the barrier profile. To reduce this trapping effect we tunnel through high quality gate oxide, which has far less trapping than interpoly oxide. Both injection and tunneling have very stable and repeatable characteristics. When tunneling at a fixed oxide voltage, the tunneling current decreases only 50 percent after $10nC$ of charge has passed through the oxide. This quantity of charge is orders of magnitude more than we would expect a synapse to experience over a lifetme of operation.

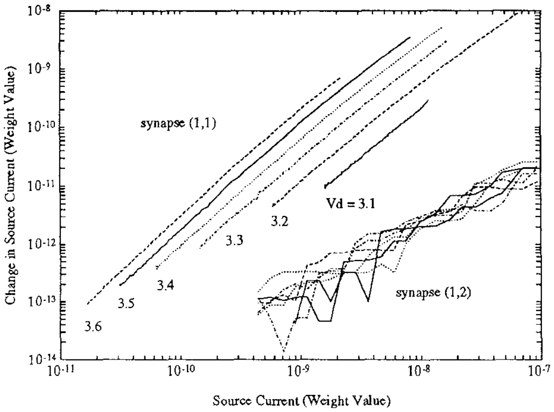

Figure 3: Source Current Decrement during injection vs. Source Current for several values of drain voltage. The injection operation decreases the synaptic weight. $Vg2$ was held at $0V$, and $Vg1$ was at $8V$ during the $0.5s$ injecting pulses. The change in source current is approximately proportional to the source current to the β power, where of β is between 1.7 and 1.85 for the range of drain voltages shown. The change in source current in synapse (1,2) is much less than the corresponding change in synapse (1,1) and is nearly independant of drain voltage. The effect of this injection on synapses (2,1) and (2,2) is negligible.

Figure 4 shows measured data on the change in source current during tunneling as a function of source current for several values of tunneling voltage. We can predict the functional form of the curves in Fig. 4. From [2], the functional form of tunneling current is

$$I_{fg} = I_{o_{tun}} e^{-\frac{V_o}{V_{tun}-V_{fg}}} \tag{7}$$

By expanding V_{fg} for fixed V_{tun} as $V_{fg0} + \Delta V_{fg}$ and inserting (1), we get

$$I_{fg} = I_{tun}(V_{fg0}) \left(\frac{I}{I_{fg0}}\right)^{\alpha-1} \tag{8}$$

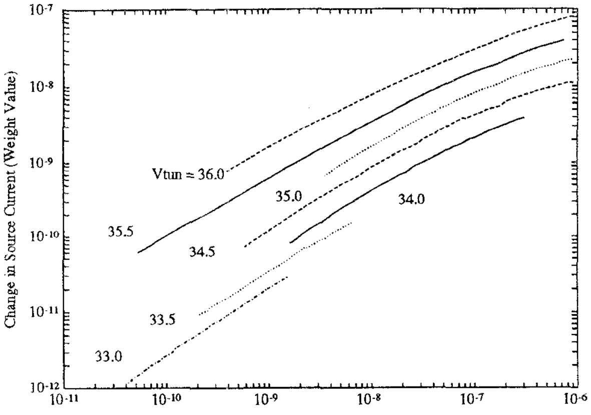

Figure 4: Synapse (1,1) source current increment vs. Source Current for several values of tunneling voltage. The tunneling operation increases the synaptic weight. $Vg1$ was held at $0V$ and $Vg2$ was $8V$ while the tunneling line was pulsed for $0.5s$ from $20V$ to the voltage shown. The change in source current is approximately proportional to the α power of the source current where α is between 0.7 and 0.9 for the range of tunneling voltages shown. The effect of this tunneling proceedure on synapse (2,1) and (2,2) are negligible. Selectivity among synapses on the same row is shown in Fig. 5.

Tunneling removes electrons from the floating gate; therefore the floating gate current is positive. The resulting current change is

$$\frac{dI}{dt} = \frac{I_{tun}(V_{fg0})}{Q_o} I_{fg0} \left(\frac{I}{I_{fg0}} \right)^{\alpha} \tag{9}$$

The model qualitatively agrees with the data in Fig. 4, with α in the range of 0.7 - 0.9. The tunneling selectivity between synapses on different rows is very good, but tunneling selectivity along along a row is poor. We typically measuere tunneling selectivity ratios along a row between 3 - 7 for our devices.

V. Conclusion - Model of the Array of Single Transistor Synapses

wE have designed, fabricated, characterized, and modeled an array of single transistor synapses. This single transistor synapse simultaneously performs long term weight storage, the product of the input and floating gate voltage, and increases and decreases the weight accorind to a hebbian or backpropagation learning rule. We have shown that the injection process is very selective along a row or column. We have found that the tunneling process is only weakly selective for a particular synapse along a row. More work is required to improve the selectivity of tunneling along a row. Using the steady state source current as the representation of the weight value, both the incrementing and decrementing functions are proportional to a power of the source current.

Parameter	Typical Values	Parameter	Typical Values
β	1.7 - 1.9	α	0.7 - 0.9
γ_d	0.030	γ_{tun}	0.027
γ_g	0.25	δ	0.02 - 0.1
Q_o	$.2\ pC$	V_{inj}	$78mV$

Table 1: Typical measured values of the parameters in the modeling of the single transistor synapse array.

Finally, we present an approximate model of our Array of these Single Transistor Synapses. The learning increment of the synapse at position (i,j) can be modeled as

$$\begin{aligned} \frac{dI_{ij}}{dt} \quad = \quad & \frac{1}{Q_o} \left(I_{tun}(V_{fg0}) I_{fg0} \left(\frac{I}{I_{fg0}} \right)^{\alpha} - A I_{ij}^{\beta} e^{\frac{V_d}{V_{inj}}} \right) \\ & + \frac{I_{ij}}{Q_o} \left(\gamma_d \frac{dV_{d_i}}{dt} + \gamma_{tun} \frac{dV_{tun_i}}{dt} + \gamma_g \frac{dV_{g_j}}{dt} \right) \end{aligned} \tag{10}$$

for the synapse at position (i,j), where γ_d is the drain to floating gate coupling coefficient, γ_{tun} is the tunneling line to floating gate coupling coefficient, and γ_g is the gate to floating gate coupling coefficient. Typical values for the parameters in (10) are given in Table 1.

VI. Acknowledgements

The work was supported by the office of Naval Research, the Advanced Research Projects Ageny, and the Beckman Foundation.

References

[1] P. Hasler, C. Diorio, B. A. Minch, and C. Mead, "Single Transistor Learning Synapses", *Advances in Neural Information Processing Systems 7, Morgan Kaufmann Publishers, San Mateo, CA, 1995.* M. Holler, S. Tam, H Castro, R. Benson,

[2] M. Lenzlinger and E. H. Snow, "Fowler - Nordheim Tunneling into Thermally Grown SiO_2", *Journal of Applied Physics, Vol. 40, NO. 1 pp. 278 - 283.*

[3] P. Hasler, C. Diorio, B. A. Minch, and C. Mead, "An Analytic model of Hot Electron Injection from Boltzman Transport", *Tech. Report 123456*