A Neuron-MOS Neural Network Using Self-Learning-Compatible Synapse Circuits

Tadashi Shibata, *Member, IEEE,* Hideo Kosaka, Hiroshi Ishii, *Member, IEEE,* and Tadahiro Ohmi, *Member, IEEE*

***Abstract*—A circuit technology for self-learning neural network hardware has been developed using a high-functionality device called Neuron MOS Transistor (νMOS) as a key circuit element. A νMOS can perform weighted summation of multiple input signals and thresholding all at a single transistor level based the charge sharing among multiple capacitors. An electronic synapse cell has been constructed with six transistors by merging a floating-gate EEPROM memory cell into a new-concept νMOS differential-source-follower circuitry. The synapse can represent both positive (excitatory) and negative (inhibitory) weights under single V_{DD} power supply and is free from standby power dissipation. An excellent linearity in the weight updating characteristics of the synapse memory has been also established by employing a simple self-feedback regime in each cell circuitry, thus making it fully compatible to the on-chip self-learning architecture of νMOS neural networks. The basic operation of the synapse cell and a νMOS neural network using the synapse has been experimentally verified using test circuits fabricated by a double-polysilicon CMOS process.**

I. Introduction

NEURAL NETWORKS are now drawing considerable attention as a new paradigm of information processing because of their self-adaptive problem-solving capabilities [1]–[3]. A number of neural network algorithms have been proposed, but their implementation was mostly in software programs running on digital computers. In order to explore real-world applications of neural networks, however, their hardware implementation on silicon chips [3], [4], with high integration density and on-chip self-learning capability, is critically demanded.

Although the implementation using digital circuit technology [5]–[9] is superior in terms of high-precision computation, a huge number of transistors are required to build multiply-accumulate units as well as synaptic weight memories. In order to achieve ultralarge-scale integration of neural networks, analog circuit implementations [10]–[14] are preferred because of their much simpler circuit configurations.

In some of the analog approaches current-mode circuitries are employed for basic neuron computations. Namely, the signal summation is conducted by the wired-sum technique based on the Kirchhoff's current law, and the multiplication at synapses is also carried out in current mode [10] using the four-quadrant MOS analog multiplier [15]. Such current-mode computations, however, allow dc currents to flow, resulting in a steady state power dissipation on a chip. This would present a difficult issue for ultralarge scale integration (ULSI) of analog neural networks because reducing the total power dissipation on a chip is one of the most critical issues of ULSI systems. Exception is the inherently low-power subthreshold-logic approach in which currents flowing in MOSFET's in the subthreshold regime are utilized for computation [1].

We have explored an unique physical computing scheme employing capacitance coupling phenomena in which the summation of voltage signals is carried out by charge sharing among multiple capacitors. Therefore the sum operation itself is in principle power-dissipation free. This linear summing operation has been integrated into a MOSFET structure and, as being combined with the transistor's thresholding action, a neuron-like functionality has been created at a single transistor level (Fig. 1). The new functional device has been named Neuron MOSFET (neuMOS or νMOS in short) [16], [17] due to its functional similarity to the mathematical model of a neuron [18], and has been applied to build a number of new-architecture logic integrated circuits [19]–[23]. The purpose of this paper is to present the circuit technology to build analog neural networks having on-chip self-learning capability using Neuron MOSFET as a key circuit element.

The νMOS neural network has such an unique architecture that it works purely in the voltage mode of operation. Namely, the quantities such as neuron output signals and synaptic weights are all represented by analog voltage values and their multiplication and summation are all carried out in the voltage signal domain.

For synaptic weight memories, we employed a floating-gate EEPROM technology as is generally the case [24]–[27]. However, by merging an EEPROM memory cell into a νMOS differential source follower circuitry, such a unique feature of an electronic synapse has been established that it is free from standby-power dissipation and capable of representing both positive (excitatory) and negative (inhibitory) weights under single V_{DD} power supply. Such a synapse circuit has been composed of only six transistors. How to achieve a good linearity in the data updating characteristics of a floating gate EEPROM memory, on the other hand, is one of the most difficult issues of its synapse application, and several techniques have been proposed for the device structure [28], [29] as well as for the circuit architecture [10]. An excellent linearity under constant programming pulses has been achieved for the first time [30] in such a simple manner as just adding one more transistor to the six-transistor synapse circuit.

Manuscript received March 19, 1993; revised March 7, 1995. This work was supported in part by the Ministry of Education, Science and Culture of Japan, under Grant-in-Aid for Developmental Scientific Research 05505003.

The authors are with the Department of Electronic Engineering, Tohoku University, Aoba-ku, Sendai 980-77 Japan.

IEEE Log Number 9412843.

0018–9200/95$04.00 © 1995 IEEE

Authorized licensed use limited to: TOHOKU UNIVERSITY. Downloaded on March 01,2010 at 23:44:54 EST from IEEE Xplore. Restrictions apply.

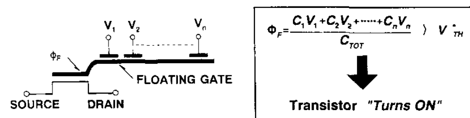


Fig. 1. Symbolic representation of neuron MOS transistor (νMOS).

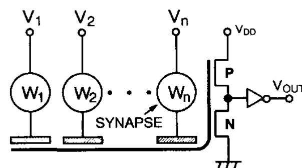


Fig. 2. Schematic of a single neuron module composed of a complementary νMOS inverter and synapse cells. The coupling capacitors are made all identical.

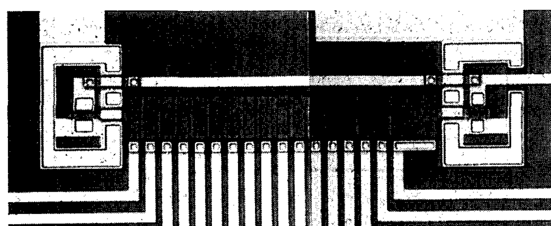
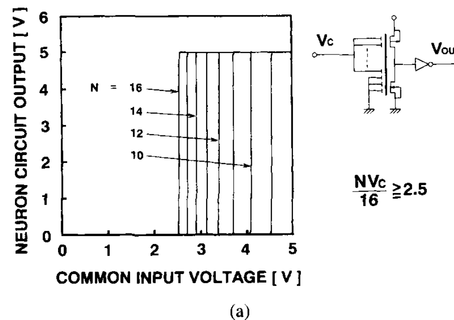
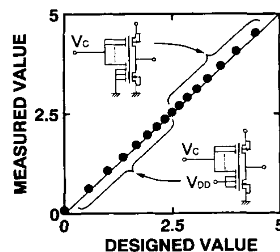


Fig. 3. Photomicrograph of 16-input complementary νMOS inverter fabricated by a double-polysilicon CMOS process.



(a)



(b)

Fig. 4. (a) Output characteristics of 16-input neuron cell shown in Fig. 3. The neuron gets fired when $NV_C/16 \geq V_{\rm INV}^{*}(= 2.5$ V). (b) Comparison between the measured and designed values of apparent thresholds of the neuron cell.

The organization of this article is as follows. Firstly the basic neuron module is presented in Section II. Then in Section III we will describe the principle and operation of the six-transistor synapse cell in detail. The operation of the synapse circuit in a νMOS neural network hardware is also demonstrated by experiments. In Section IV the experimental data for the weight updating characteristics of synapse memories are presented. And it will be shown that the characteristics can be linearized by a simple modification in the original six-transistor cell circuitry. Then the concluding remarks are given in Section V.

II. Neuron-Cell Circuitry

A single neuron module used in νMOS neural networks is schematically shown in Fig. 2, where a neuron is composed of a complementary νMOS (C-νMOS) inverter having all identical coupling capacitances and a regular CMOS inverter. $V_1 \sim V_n$ are the outputs of previous-layer neurons, being multiplied by the respective weights at synapses, and then, transferred to the neuron inputs. The net $= \sum W_i V_i$ is automatically calculated by charge redistribution on the common gate of a C-νMOS inverter which is electrically floating, and then squashed into 0 V or $V_{\rm DD}$ by two-stage CMOS inverter action. Therefore the common floating gate of the C-νMOS inverter acts as a dendrite of a neuron. The net charge on the common gate in a floating state Q_F is made constant for the neuron cell application. (Q_F is usually treated as 0 in logic circuit applications [20], which is valid under thermal equilibrium and is easily achieved by UV erasing after chip fabrication (see Fig. 7 in [20]) or by periodic refreshing of the charge on the floating node through a clock-driven switching transistor connected to the common gate [41]. An example for such a scheme is given in Fig. 9(b) of the present article.) For the synapse cell application, on the other hand, the charge on a floating gate is taken as a variable to represent the synaptic weight, which will be described in detail in Sections III and IV.

Fig. 3 shows the photomicrograph a C-νMOS inverter having 16 input terminals of identical coupling capacitors, which was fabricated by a standard double-polysilicon CMOS process. The circuit was used to verify the principal functions of a neuron cell, i.e., the summation and thresholding. The measured data are presented in Fig. 4(a) and (b). In Fig. 4(a), some of the input gates (N indicates the number of such gates) are connected together and are given with a common input voltage changing from 0 to 5 V. The rest of the gates are all grounded. The threshold at which the neuron fires is increasing as N increases. This apparent threshold voltage is shown in Fig. 4(b) as a measured-value versus designed-value plot. The upper half of the data are those from Fig. 4(a) and the lower half was obtained by giving 5 V to the rest of the gates. Excellent agreement is seen, indicating the neuron-cell in Fig. 3 functions exactly as expected.

It should be commented here that the C-νMOS inverter in the scheme of Fig. 2 is biased in the transition region where a dc current flows, leading to the increased power dissipation. In order to reduce the power, we are proposing a scheme in which the floating gate is equipped with a clock-driven switch to cut off such a dc current except for the period when neurons are being activated [41].

Authorized licensed use limited to: TOHOKU UNIVERSITY. Downloaded on March 01,2010 at 23:44:54 EST from IEEE Xplore. Restrictions apply.

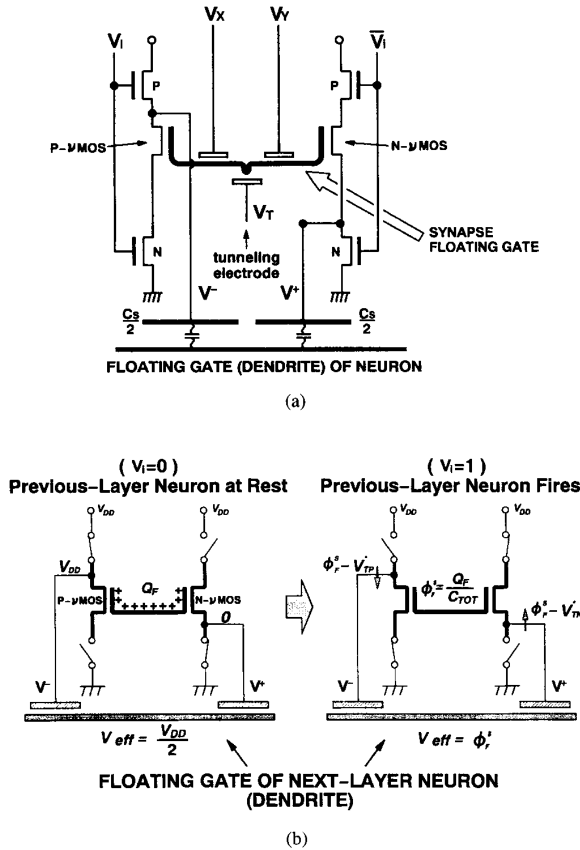


Fig. 5. (a) νMOS differential source follower circuitry for an electronic synapse memory cell which is composed of six transistors. (b) Switching states of CMOS inverters in the cell when the previous-layer neuron is resting (left) or firing (right).

III. νMOS Differential Source-Follower Synapse Cell

As is clearly seen in Fig. 4(a), our neuron cell has a hard-limiting characteristics of a step function, *viz.*, its output is either a binary 0 or a binary 1. This simplifies the weight multiplication operation conducted by a synapse cell in the scheme of Fig. 2. What is required for a synapse cell to carry out is firstly to memorize an analog weight data W_i, and then to transfer the weight W_i to the next-layer neuron when $V_i =$ "1," i.e., when the previous-layer neuron fires. When the previous-layer neuron is resting and $V_i = 0$, the synapse gives a neutral output voltage, which is $V_{\text{DD}}/2$ in the present work.

A. Six-Transistor EEPROM Synapse Cell

The circuit diagram of the six-transistor synapse cell is given in Fig. 5(a). The cell is basically a floating-gate EEPROM memory and the charge on the floating gate (FG) represents the weight value of the synapse. Both N-channel νMOS and P-channel νMOS read out the amount of the FG charge non-destructively through the source follower action, and transfer the data as voltage signals to the common gate (dendrite) of a next-layer neuron via two coupling capacitors of identical magnitude $(C_S/2)$.

The right-hand side of the circuit is an N-νMOS source follower which is merged into a regular CMOS inverter to cut off the dc current path. The left-hand side is a P-νMOS source follower also merged into a CMOS inverter, thus achieving the standby-power free feature of the cell. The output signal from a previous-layer neuron, V_i, controls the on and off states of the four transistors in the CMOS inverters. The switching states of the cell depending on whether $V_i = 0$ or 1 are depicted in Fig. 5(b).

Both the N-νMOS and P-νMOS share the common floating gate (FG) and the dual equally-weighted input gates, V_X and V_Y, controls the potential of the synapse floating gate ϕ_F^S for programming. The tunneling electrode is separated from the FG by an intervening very thin SiO_2 film (~100 Å thick) to allow charge transfer between the tunneling electrode and the FG when high-voltage programming pulses are applied to V_X and V_Y. The net charge on the synapse floating gate Q_F^S represents the synaptic weight, and the value is updated by charge injection/extraction during programming. In the following the data readout operation of the cell is firstly explained.

B. Data Readout/Transfer Operation

The outputs of the source-followers are both connected to the common gate (dendrite) of a next-layer neuron via an identical coupling capacitance of $C_S/2$. As a result, these two output voltages are intermixed and averaged at the electrically-floating common gate, and the net result reduces to giving an effective input voltage $V_{\text{eff}} = (V^- + V^+)/2$ to the dendrite through a single capacitor of C_S. Fig. 5(b) illustrates the switching states of the CMOS inverters in the cell.

When the previous-layer neuron is at rest $(V_i = 0)$, the output nodes of N- and P-νMOS′ are precharged to 0 and V_{DD}, respectively. Therefore, the effective output of the synapse cell becomes $V_{\text{eff}} = V_{\text{DD}}/2$. This value is taken as a level of reference, which corresponds to the neutral output of a synapse. When the previous-layer neuron fires, both source followers are activated and the output node of the N-νMOS source follower (V^+) approaches $\phi_F^S - V_{\text{Tn}}^*$, while that of the P-νMOS (V^-) approaches $\phi_F^S - V_{\text{Tp}}^*$. Here, ϕ_F^S, V_{Tn}^*, and V_{Tp}^* denote the potential of the common floating gate and the threshold voltages of the N-νMOS and P-νMOS as seen from the common floating gate, respectively. The effective input to the dendrite in this case becomes $V_{\text{eff}} = \phi_F^S$, provided the N- and P-νMOS′ have the same thresholds of opposite signs. The value of ϕ_F^S is determined as Q_F^S/C_{TOT}^S, where Q_F^S and C_{TOT}^S are the total charge and total capacitance of the common floating gate of the synapse circuit. (Superscript S was given to remind the quantity is related to a synapse circuit.) Therefore the value of ϕ_F^S, accordingly the synaptic weight, can be programmed as the total amount of the floating gate charge Q_F^S.

The time variations of the output voltages of N- and P-νMOS source followers as calculated by HSPICE simulation are shown in Fig. 6 for cases in which N-νMOS and P-νMOS are both enhancement-mode transistors (Fig. 6(a) and (b)) or both depletion-mode transistors (Fig. 6(c) and (d)) . The

Authorized licensed use limited to: TOHOKU UNIVERSITY. Downloaded on March 01,2010 at 23:44:54 EST from IEEE Xplore. Restrictions apply.

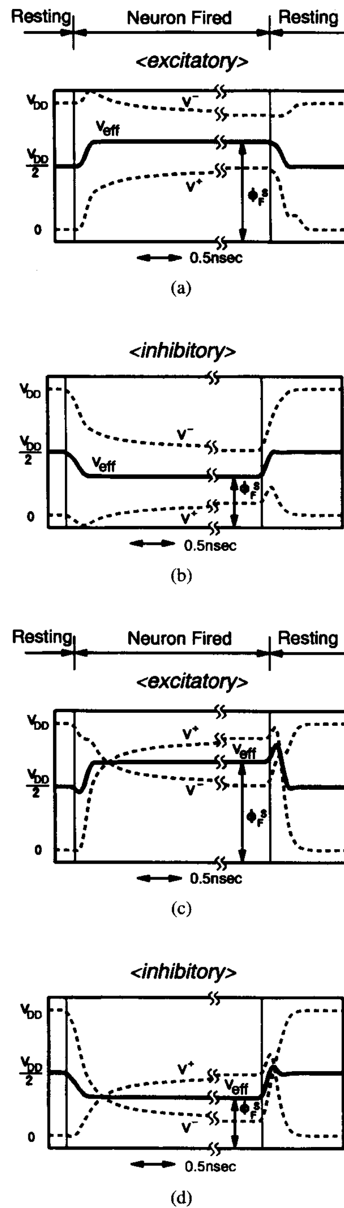


Fig. 6. Output waveforms of N-νMOS (V^+) and P-νMOS (V^-) source followers in the synapse circuit of Fig. 5(a) and their averages ($V_{\text{eff}} = (V^+ + V^-)/2$) as calculated by HSPICE simulation. Here V_i changes from 0 to V_{DD} (5 V) and back to 0 with the rise and fall times of 0.2 ns. In (a) and (b), N-νMOS and P-νMOS are both enhancement-mode transistors, while they are both depletion-mode transistors in (c) and (d). The absolute values for thresholds are all 1 V.

substrate doping concentrations were $N_A = N_D = 2 \times 10^{14}$ cm^{-3}, and the absolute values of all threshold voltages were set at 1 V. Here C_S was chosen as 20 fF and the dendrite potential (the floating potential of the common gate of the next-layer neuron, so called hereafter to avoid confusion with ϕ_F^S) is assumed for the moment to be zero and not changing. The average of V^+ and V^- indicated as V_{eff} gives an effective synapse output voltage to the dendrite. When V_{eff} is greater than the reference level $V_{\text{DD}}/2$, the circuit represents an excitatory synapse. If V_{eff} is smaller than $V_{\text{DD}}/2$, the circuit becomes an inhibitory synapse. Therefore Fig. 6(a) and (c) demonstrates the behavior of excitatory synapses (ϕ_F^S = 3.5 V), and Fig. 6(b) and (d) those of inhibitory synapses (ϕ_F^S = 1.5 V).

Charging up a capacitor load by a source follower is a relatively slow process because the transistor tends to cut off when the output approaches its destination. However, it is quite interesting to observe that V_{eff} saturates to the final value much faster than V^+ or V^- does due to the cancellation effect of the two signals changing in a nearly anti-symmetrical manner. Therefore the inherently slow operation of source-followers has been improved by the differential action of NMOS and PMOS source followers. This is a very effective way to accelerate the source follower operation.

In order to verify the operation of the differential source follower circuitry, test circuits were fabricated by a double polysilicon CMOS process (see Fig. 7(a)) and the measurement results are demonstrated in Fig. 7(b). In this particular test circuit the FG is directly connected to an external pad and ϕ_F^S is determined by an external voltage source. The measured threshold voltages were $V_{\text{Tn}}^* = -0.5$ V and $V_{\text{Tp}}^* =$ 0.2 V (both depletion mode). The measured data show similar behavior to the simulation results in Fig. 6(c) and (d). The slow operation of the test circuit is due to the stray capacitance (~several tens pF) arising from the measurement system because the circuit output was directly probed without output buffer circuitries in order to observe the unaltered characteristics.

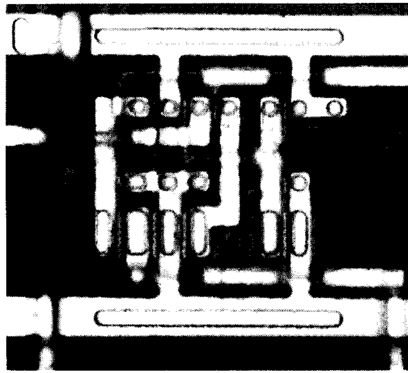
The relationship between the V_{eff} and ϕ_F^S measured for this test circuit is given in Fig. 8. The cell represents excitatory (positive-weight) or inhibitory (negative-weight) synapses depending on whether V_{eff} is greater or smaller than $V_{\text{DD}}/2$, respectively. The updating of the synaptic weight is carried out by changing the FG charge Q_F^S by programming. Q_F^S is related to the FG potential by $\phi_F^S = Q_F^S/C_{\text{TOT}}^S$.

C. Cell Programming Operation

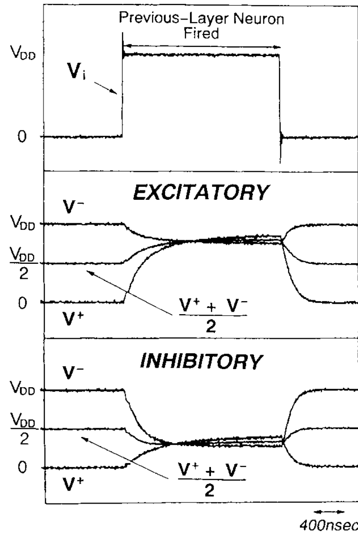
In reducing the synaptic weight, a large positive programming pulse voltage V_P is given to both V_X and V_Y while keeping the tunneling-electrode grounded ($V_T = 0$). Then the FG potential is pulled up and it becomes $\phi_F^S \approx V_P$ provided C_P (the coupling capacitance of V_X or V_Y terminal) is the dominant contribution to the total synapse floating-gate capacitance C_{TOT}^S. Due to the large electric field in the tunneling oxide, electrons are injected into the floating gate via Fowler–Nordheim tunneling and negative incremental charge is added to Q_F^S, thus reducing the weight. The increase in the synaptic weight is carried out similarly by reversing the voltage polarity.

If either one of V_X or V_Y is grounded, the tunnel oxide field is reduced by a factor of approximately 1/2 because $\phi_F^S \approx V_P/2$, resulting in a drastic reduction in the tunneling current by a factor of about $\sim 10^{-11}$ due to the exponential electric-field dependence of the Fowler–Nordheim current [32]. Therefore data updating occurs selectively only at cells in

Authorized licensed use limited to: TOHOKU UNIVERSITY. Downloaded on March 01,2010 at 23:44:54 EST from IEEE Xplore. Restrictions apply.



(a)



(b)

Fig. 7. (a) Photomicrograph of a test synapse cell circuit and (b) measured output waveforms of N-νMOS (V^+) and P-νMOS (V^-) source followers. Their average (calculated on a digitizing oscilloscope) represents the effective output of the synapse, i.e., the weight value. The stray capacitance arising from the direct probing is a few thousands times larger than C_S. Here $V_{DD} = 5$ V and ϕ_F^S values were set to 3 V and 0 V for excitatory and inhibitory synapses, respectively.

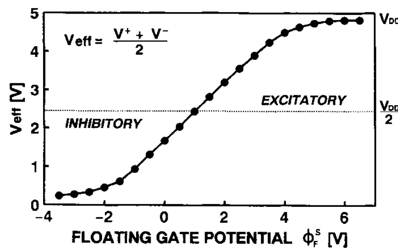
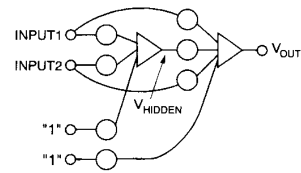
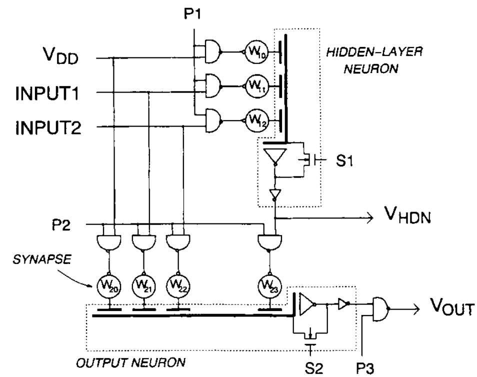


Fig. 8. Measured relationship between effective synapse output V_{eff} and synapse floating-gate potential ϕ_F^S. In this measurement, the ϕ_F^S voltage was directly supplied from an external voltage source.

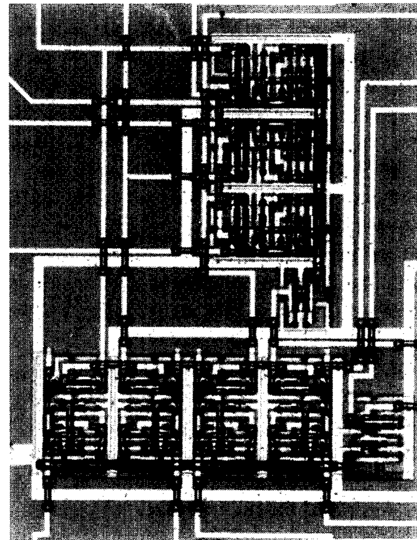
which both V_X and V_Y are pulled up to V_P. If V_X and V_Y are connected to interconnections running x and y directions over



(a)



(b)



(c)

Fig. 9. Test circuit of νMOS neural network. (a) Schematic of network configuration. (b) Circuit diagram. (c) Photomicrograph of a test circuit fabricated by a double polysilicon CMOS process.

the synapse cell array matrix, the weight modification occurs only at the crossing points of particular V_X and V_Y lines given with a high programming pulse. Such characteristics are very conveniently utilized in implementing Hebbean-like learning algorithms directly on the hardware [33]. The selective cell programming scheme presented here was first introduced in the Dual Control-gate EEPROM cell (dc cell) [34].

During the normal operating mode of a neural network, V_X, V_Y and V_T are all set to 0 V. Then the synapse FG

Authorized licensed use limited to: TOHOKU UNIVERSITY. Downloaded on March 01,2010 at 23:44:54 EST from IEEE Xplore. Restrictions apply.

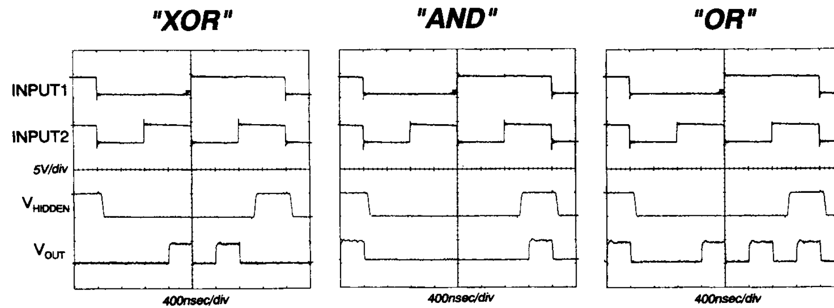


Fig. 10. Measured output characteristics of the test νMOS neural network which learned XOR, AND, and OR functions.

potential becomes $\phi_F^S = Q_F^S/C_{\rm TOT}^S$, which represents the synaptic weight strength according to the relation like the one shown in Fig. 8.

D. νMOS Neural Network Operation

In order to verify the operation of a neural network utilizing the νMOS neuron cell and the synapse cell, a simple test circuit was designed. Fig. 9(a) and (b) show the network configuration and the circuit diagram, respectively. It consists of one output-layer neuron, one hidden-layer neuron, and five synapses with two input-to-output jumping connections. Two more synapses are provided for threshold adjustment of neurons with constant $V_{\rm DD}$ inputs. $P_1 \sim P_3$ are enabling clock signals for synchronous forward operation. When $P_2 = 0$, for instance, the outputs of synapses $W_{20} \sim W_{23}$ are all reset to the neutral level of $V_{\rm DD}/2$. Under this condition, the signal S_2 short-circuits the common gate and the output of the C-νMOS inverter each time before the neuron operates and auto-adjusts its inverting threshold. This is the well-known offset cancellation technique employed in comparators for A/D converters [35] and would help to enhance the accuracy of νMOS inverter operation in large systems.

In this test circuit, the synapse cell shown in Fig. 7(a) was employed and the weight values were defined by external voltage sources. In order to determine the weight values of synapses, computer simulation was conducted using Hardware Backpropagation (HBP) learning algorithm [36]–[38], which is a simplified version of the original Backpropagation algorithm [39]. We have developed HBP aiming at facilitating the hardware implementation of a learning algorithm on νMOS neural networks. The network successfully learned the XOR function and the weight values determined by the simulation were $W_{10} = 3$ V; $W_{11} = W_{12} = 2.6$ V; $W_{20} = 1$ V; $W_{21} = W_{22} = 3.2$ V; $W_{23} = -1.6$ V.

Fig. 9(c) shows the photomicrograph of a test circuit fabricated by a double-polysilicon CMOS process and the measurement results on the test circuit whose synaptic weights were determined as mentioned above are shown in Fig. 10. The circuit exhibits the right response of an XOR function. The learned states of the test circuit of AND and OR functions are also shown in the Fig. 10. In this way the basic operation of a νMOS neural network using the νMOS differential-source-follower synapse circuits has been experimentally verified.

E. Bootstrap Effect

One potential problem arising from the circuit configuration in which a number of source follower circuits are interacting through capacitance coupling is addressed and discussed in the following.

In calculating the characteristics in Fig. 6, the dendrite potential was assumed not changing and always grounded. However, this is not the case because the dendrite potential does change through the capacitance coupling with many source followers. Then the change in the dendrite potential $\Delta\phi_F^D$ is fed back to the V^+ or V^- node of each source follower through the coupling capacitor of $C_S/2$, and can affect its operation.

If the output node of an N-νMOS source follower, for instance, is pushed up by this effect and exceeds its final value $V^+(t=\infty) = \phi_F^S - V_{\rm Tn}^*$, the transistor is turned off because the FG-to-source bias of the N-νMOS becomes smaller than the threshold. Then the source follower is disconnected from the dendrite. This is a well recognized effect and is utilized in the bootstrap pull-up inverter circuitry. It is anticipated that this would cause a problem in the operation of a neuron cell. In the following we discuss this bootstrap effect, and it is shown that the problem can be reasonably resolved or even favorably utilized in enhancing the noise margin of neuron operation.

In order to examine the influence of the boot-strap effect, we consider the problem taking an N-νMOS source follower operation as an example. Since our νMOS neural network is designed to operate under a system clock, all circuits operate synchronously. Before synapse circuits are activated, the dendrite potential is at $V_{\rm DD}/2$. If the dendrite potential is decreasing after synapses are activated, the disconnection due to the bootstrap effect never happens in N-νMOS$'$. If the dendrite potential is increasing, however, the disconnection can occur under certain circumstances. Now let us find the condition to avoid the disconnection by semi-qualitative consideration.

To make it simple, assume for the moment that the time change of the dendrite potential is much faster than the time change of the output voltage of an N-νMOS source follower. Then the increase in the dendrite potential $\Delta\phi_F^D$ appears instantaneously at the V^+ node by almost the same amount, and V^+ is suddenly raised from 0 V to $\Delta\phi_F^D$. In order to prevent the N-νMOS from being disconnected by

Authorized licensed use limited to: TOHOKU UNIVERSITY. Downloaded on March 01,2010 at 23:44:54 EST from IEEE Xplore. Restrictions apply.

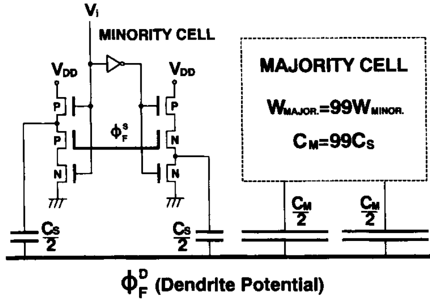


Fig. 11. Test circuit configuration utilized for analyzing the bootstrap effect by HSPICE. 100 synapse cells are connected to the dendrite of a single neuron cell, where the majority cell represents 99 cells and the minority cell represents a single cell.

bootstrapping, its final value $V^+(t=\infty)$ must be larger than $\Delta\phi_F^D$. Since the maximum of $\Delta\phi_F^D$ is $V_{\rm DD}/2$ (because the initial value of ϕ_F^D is $V_{\rm DD}/2$ and its maximum value is $V_{\rm DD}$), it is requested that $V^+(t=\infty) \geq V_{\rm DD}/2$. Therefore, the minimum allowable value for $V^+(\infty)$ without accompanying the bootstrap effect would become $V_{\rm DD}/2$. The effective synapse output, i.e., the weight is given by $(V^+(t=\infty) + V^-(t=\infty))/2$. Therefore the minimum weight free from the bootstrapping in any occasion is obtained by setting both $V^+(t=\infty)$ and $V^-(t=\infty)$ at their minimum values of $V_{\rm DD}/2$ and 0, respectively, yielding the value of $V_{\rm DD}/4$. In order to represent this minimum weight, we need to set $\phi_F^S = V_{\rm DD}/4$, and at the same time it is necessary to specify the depletion-mode thresholds of N-νMOS and P-νMOS at $-V_{\rm DD}/4$ and $V_{\rm DD}/4$, respectively. Similar considerations for a P-νMOS source follower yields the upper limit for the weight of $(3/4)V_{\rm DD}$, assuming the same depletion-mode thresholds for N-νMOS and P-νMOS.

Then we may draw a conclusion that the bootstrapping can be avoided as long as $V_{\rm DD}/4 < V_{\rm eff} < (3/4)V_{\rm DD}$ and both N- and P-νMOS′ have depletion-mode thresholds whose absolute values are $V_{\rm DD}/4$. The bootstrapping can occur when the weight becomes smaller than $V_{\rm DD}/4$ or larger than $(3/4)V_{\rm DD}$, but it does not happen always. The bootstrapping only occurs for a synapse whose output goes to an extreme end opposite to the direction of the majority. In other words, the opinions of minority synapses are disregarded when they are in an extreme opposition to those of the majority. Since the decision of "0" or "1" based on the average of all synapse outputs is the very function of a neuron, this bootstrap effect can contribute to enhancing the noise margin of the neuron operation.

The above semi-qualitative discussion is verified by HSPICE simulation in the following. Simulations were conducted on a test circuit shown in Fig. 11 where the dendrite potential ϕ_F^D is determined by two synapses, i.e., a minority cell and a majority cell. The minority cell represents a single synapse and the majority represents the all other synapses connected to the same dendrite in which all transistors' channel width and the coupling capacitors were multiplied by a factor 99. We tried to envisage what happens to the output of a 1% minority synapse trying to pull down the dendrite potential in opposition to the 99% of majority trying to pull up the dendrite potential.

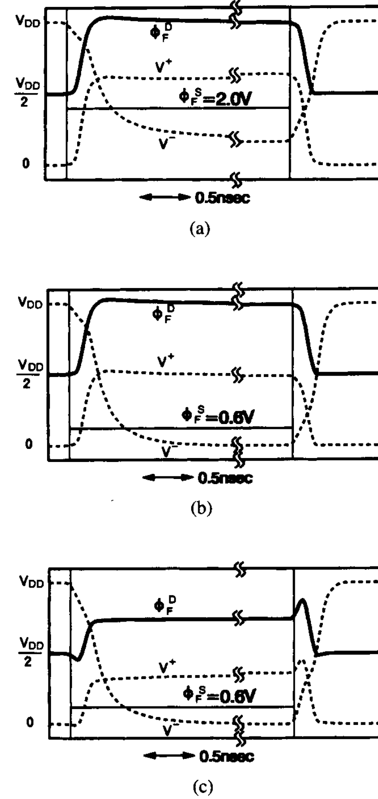


Fig. 12. Results of HSPICE simulation demonstrating the bootstrap effects occurring in the circuit shown in Fig. 11. The output of the majority cell is 5 V for (a) and (b) and 3.75 V for (c). The bootstrapping occurs in (b).

The results of the worst case simulations are demonstrated in Fig. 12(a) and (b) in which the dendrite potential ϕ_F^D goes to the maximum of $V_{\rm DD}$ (5 V). (For this end ϕ_F^S of the majority cell was set at 6.25 V to give the effective synapse output of 5 V. The substrate doping concentrations of 2×10^{14} cm^{-3} were utilized to minimize the body effect. Depletion thresholds of $V_{\rm Tn}^* = -1.25$ V $(-V_{\rm DD}/4)$, $V_{\rm Tp}^* = 1.25$ V $(V_{\rm DD}/4)$ were employed.) If $\phi_F^S = 2.0$ V (weight $= 2.0 > V_{\rm DD}/4$) in the minority cell, bootstrapping disconnection does not occur as shown in Fig. 12(a). If $\phi_F^S = 0.6$ V (weight $= 0.6 < V_{\rm DD}/4$) in the minority cell, however, the bootstrapping disconnection is observed in the N-νMOS source follower because its output V^+ is seen to saturate to a value (2.4 V) higher than its anticipated final value of $\phi_F^S - V_{\rm Tn}^* (= 1.85$ V) as shown in Fig. 12(b). On the other hand, however, if the dendrite potential goes to 3.75 V, the bootstrapping does not occur as shown in Fig. 12(c).

In Fig. 13 we summarize the results of simulation by illustrating the regions where the bootstrap disconnection occurs. It happens only in limited regions where the output of a minority cell goes to an extreme end opposite to the movement of the majority. The boundaries predicted by the preceding semi-qualitative analysis is also shown by broken lines in the figure, demonstrating the prediction is quite reasonable.

Authorized licensed use limited to: TOHOKU UNIVERSITY. Downloaded on March 01,2010 at 23:44:54 EST from IEEE Xplore. Restrictions apply.

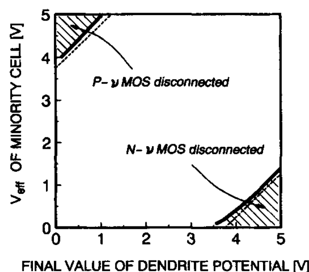


Fig. 13. Hatched regions in the figure indicate where the disconnection of N-νMOS or P-νMOS in the minority cell occurs for various combinations of $V_{\rm eff}$ of the minority cell and the final value of the dendrite potential. Bold lines were determined by HSPICE simulation using the test circuit given in Fig. 11, while the broken lines indicate the results of the semi-qualitative analysis developed in the text. Here, $V_{\rm Tn}^{*} = -1.25{\rm V}(-{\rm V_{DD}}/4)$ and $V_{\rm Tp}^{*} = 1.25{\rm V}({\rm V_{DD}}/4)$ and the substrate doping of 2×10^{14} cm^{-3} was assumed for both P and N substrates.

IV. Weight-Updating Characteristics

In order to demonstrate the cell programming characteristics by experiment, the apparent threshold voltage $V_{\rm TH}$ of N-νMOS in Fig. 5(a) was monitored using both V_X and V_Y terminals as a gate electrode, because ϕ_F^S cannot be directly measured. The $V_{\rm TH}$ value was plotted in Fig. 14 as a function of the number of programming pulses. The programming was performed by giving high voltage pulses to both V_X and V_Y terminals. $V_{\rm TH}$ changes by a large amount by a single pulse, followed by a very gradual increase up to 50 pulses, thus showing a strong nonlinearity. The reduction in the pulse width does not improve the nonlinearity but only the final saturation value is reduced.

EEPROM cell writing by applying constant voltage pulses exhibits such a strong nonlinear dependence on the number of pulses due to the exponential dependence of the Fowler–Nordheim tunneling current on the electric field in the tunnel oxide. Once electrons are injected to the FG, the FG potential is reduced by the increased negative charge, resulting in the suppression of further charge injection. Such nonlinear characteristics severely degrades the learning performance of a neural network, presenting one of the most critical issues of using EEPROM technology for electronic synapses [40].

As shown in Fig. 15, however, the problem has been very beautifully resolved by just including one additional NMOS transistor (Tr. 1) to the six-transistor synapse cell shown in Fig. 5(a) [30]. Before a programming pulse is applied, Tr. 1 is turned on and the output voltage of the N-νMOS source follower is set to the tunneling electrode. Then it becomes $V_T = \phi_F^S - V_{\rm Tn}^*$, and the voltage across the tunneling oxide is automatically reset to a constant value of $V_{\rm Tn}^*$ (the threshold of N-νMOS) indifferent to the amount of charge stored on the floating gate. Then a programming pulse voltage is superimposed on this constant voltage. This ensures the constant amount of charge injection (or extraction) under the same programming pulse. The details of the circuit operation are described in [42].

The weight-updating characteristics of the new seven-transistor synapse cell are shown in Fig. 16, demonstrating an excellent linearity in the threshold change. The slope of the characteristics can be altered by the pulse width as shown in the figure or by the pulse height V_P. This provides a means to control the learning rate. Such a feature is quite essential for hardware learning of neural networks. The saturation characteristics are seen at a certain level of $V_{\rm TH}$. This is because the N-νMOS source follower does not work when $\phi_F^S - V_{\rm Tn}^* \leq 0$. This is not a problem but a desirable characteristics because it prevents unnecessary charge injection to the FG. This reduces the voltage stress on the thin tunnel oxide film and enhances the long-term reliability of the synapse cell. Therefore it would be favorable in ensuring good data retention characteristics of the cell.

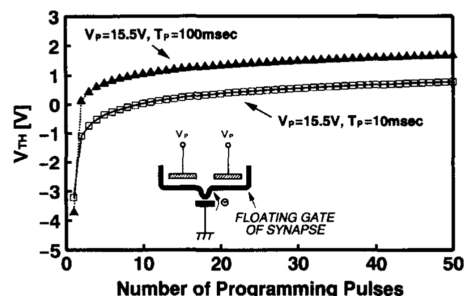
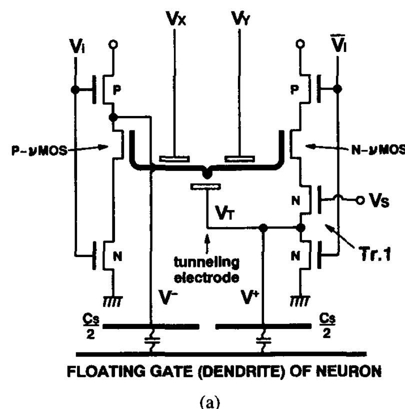
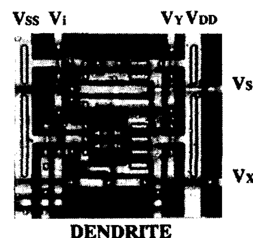


Fig. 14. Measured threshold voltage of N-νMOS in Fig. 5(a) as a function of the number of programming pulses, showing a strong nonlinearity in the programming characteristics.



(a)



(b)

Fig. 15. (a) Circuit diagram and (b) photomicrograph of improved-linearity synapse cell composed of seven transistors. One additional transistor (Tr. 1) was included in the original six-transistor cell of Fig. 5(a), which has dramatically improved the weight-updating linearity of the cell.

Authorized licensed use limited to: TOHOKU UNIVERSITY. Downloaded on March 01,2010 at 23:44:54 EST from IEEE Xplore. Restrictions apply.

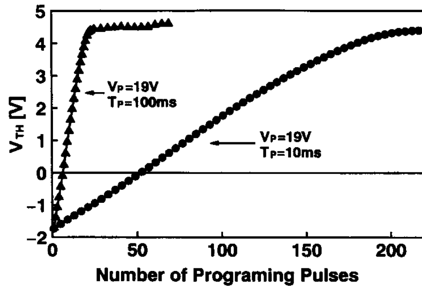


Fig. 16. Measured weight-updating characteristics of the seven-transistor synapse cell of Fig. 15, demonstrating an excellent linearity in the electron injection characteristics.

V. CONCLUSION

We have developed a circuit technology using neuron MOS transistors for building analog neural networks having on-chip self-learning capability. For this purpose, νMOS differential source follower circuitry has been developed for electronic synapse cells. The synapse memory cell composed of six transistors presents such a salient feature as standby-power free and dual-polarity representation of weights under single V_{DD} power supply. One of the most difficult issues of using floating-gate EEPROM memories for the storage of analog weight is the strong nonlinearity in the data updating characteristics. The problem has been resolved by just adding one more transistor to the original six transistor synapse-cell circuitry. The basic operation of the synapse cell and the νMOS neural network as well as the excellent weight-updating linearity of the synapse has been proven by experiments. The development of a hardware-oriented learning algorithm and the learning performance of the algorithm are discussed in [37] and [38], respectively.

The νMOS neural network architecture employing self-learning-compatible electronic synapses and its voltage mode of operation would present a unique opportunity for analog neural networks to achieve ULSI implementation on silicon chips.

ACKNOWLEDGMENT

The major part of this work was carried out in the Super Clean Room of the Laboratory for Microelectronics, Research Institute of Electrical Communication, Tohoku University.

REFERENCES

[1] C. Mead, *Analog VLSI and Neural Systems.* Reading, MA: Addison-Wesley, 1989.
[2] *Proc. IEEE*, Sept. and Oct. 1991.
[3] S. F. Zornetzer, J. L. Davis, and C. Lau, Eds., *An Introduction to Neural and Electronic Networks.* San Diego, CA: Academic, 1990.
[4] *IEEE Trans. Neural Networks*, Special Issue on Neural Network Hardware, vol. 4, no. 3, 1993.
[5] M. Yasunaga *et al.*, "Design, fabrication and evaluation of a 5-inch wafer scale neural network LSI composed of 576 digital neurons," in *Proc. IJCNN*, June 1990, pp. 527–535.
[6] M. Griffin, G. Tahara, K. Knorpp, R. Pinkahm and Riley, "An 11-million transistor neural network execution engine," in *ISSCC Dig. Tech. Papers*, Feb. 1991, TPM 11.1, pp. 180–181.
[7] B. A. White and M. I. Elmasry, "The digi-neocognitron: A digital neocognitron neural network model for VLSI," *IEEE Trans. Neural Networks*, vol. 3, no. 1, pp. 73–85, 1992.
[8] M. S. Melton, T. Phan, D. S. Reeves, and D. E. Van den Bout, "The TInMANN VLSI chip," *IEEE Trans. Neural Networks*, vol. 3, no. 3, pp. 375–383, 1992.
[9] K. Uchimura, O. Saito, and Y. Amemiya, "A high-speed digital neural network chip with low-power chain-reaction architecture," *IEEE J. Solid-State Circuits*, vol. 27, no. 12, pp. 1862–1867, 1992.
[10] M. Holler, S. Tam, H. Castro, and R. Benson, "An electrically trainable neural network chip (ETANN) with 10240 'floating gate' synapses," in *Proc. IJCNN*, June 1989, pp. 2.191–2.196.
[11] Y. Arima, K. Mashiko, K. Okada, T. Yamada, A. Maeda, H. Kondoh, and S. Kayano, "A self-learning neural network chip with 125 neurons and 10 K self-organization synapses," *IEEE J. Solid-State Circuits*, vol. 26, no. 4, pp. 607–611, 1991.
[12] B. E. Boser, E. Säckinger, J. Bromley, Y. L. Cun, and L. D. Jackel, "An analog neural network processor with programmable topology," *IEEE J. Solid-State Circuits*, vol. 26, no. 12, pp. 2017–2025, 1991.
[13] J. Van der Spiegel, P. M. Mueller, D. Blackman, P. Chance, C. Donham, R. Etienne-Cummings, and P. Kinget, "An analog neural computer with modular architecture for real-time dynamic computations," *IEEE J. Solid-State Circuits*, vol. 27, no. 1, pp. 82–92, 1992.
[14] J. A. Lanser and T. Lehmann, "An analog CMOS chip set for neural networks with arbitrary topologies," *IEEE Trans. Neural Networks*, vol. 4, no. 3, pp. 441–444, 1993.
[15] D. C. Soo and R. G. Meyer, "A four-quadrant NMOS analog Multiplier," *IEEE J. Solid-State Circuits*, vol. SC-17, no. 6, pp. 1174–1178, 1982.
[16] T. Shibata and T. Ohmi, "An intelligent MOS transistor featuring gate-level weighted sum and threshold operations," in *IEDM Tech. Dig.*, pp. 919–922, Dec. 1991.
[17] ——, "A functional MOS transistor featuring gate-level weighted sum and threshold operations," *IEEE Trans. Electron Devices*, vol. 39, no. 6, pp. 1444–1455, 1992.
[18] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biophys.*, vol. 5, pp. 115–133, 1943.
[19] T. Shibata and T. Ohmi, "Neuron MOS binary-logic integrated circuits: Part I, design fundamentals and soft-hardware-logic circuit implementation," *IEEE Trans. Electron Devices*, vol. 40, no. 3, pp. 570–576, 1993.
[20] ——, "Neuron MOS binary-logic integrated circuits: Part II, simplifying techniques of circuit configuration and their practical applications," *IEEE Trans. Electron Devices*, vol. 40, no. 5, pp. 974–979, 1993.
[21] T. Shibata, K. Kotani, and T. Ohmi, "Real-time reconfigurable logic circuits using neuron MOS transistors," in *ISSCC Dig. Tech. Papers*, Feb. 1993, FA 15.3, pp. 238–239.
[22] T. Yamashita, T. Shibata, and T. Ohmi, "Neuron MOS winner-take-all circuit and its application to associative memory," in *ISSCC Dig. Tech. Papers*, Feb. 1993, FA 15.2, pp. 236–237.
[23] R. Au, T. Yamashita, T. Shibata, and T. Ohmi, "Neuron MOS multiple-valued memory technology for intelligent data processing," *ISSCC Dig. Tech. Papers*, Feb. 1994, FA 16.3, pp. 270–271.
[24] T. Ong, P. K. Ko, and C. Hu, "The EEPROM as an analog memory device," *IEEE Trans. Electron Devices*, vol. 36, pp. 1840–1841, Sept. 1989.
[25] D. A. Durfee and F. S. Shoucair, "Comparison of floating gate neural network memory cells in standard VLSI CMOS technology," *IEEE Trans. Neural Networks*, vol. 3, no. 3, May 1992.
[26] A. Kramer, V. Hu, C. K. Sin, B. Gupta, R. Chu, and P. K. Ko, "EEPROM device as a reconfigurable analog element for neural networks," in *IEDM Tech. Dig.*, Dec. 1989, pp. 10.3.1–10.3.4.
[27] C. K. Sin, A. Kramer, V. Hu, R. Chu, and P. K. Ko, "EEPROM as an analog storage device, with particular applications in neural networks," *IEEE Trans. Electron Devices*, vol. 39, no. 6, pp. 1410–1419, June 1992.
[28] O. Fujita and Y. Amemiya, "A floating-gate analog memory device for neural networks," *IEEE Trans. Electron Devices*, vol. 40, no. 11, pp. 2029–2035, 1993.
[29] T. Kitaura, S. Sato, K. Nakajima, J. Murota, and Y. Sawada, "Switched diffusion analog memory for neural networks," in *Ext. Abst. 1993 Int. Conf. Sold State Device and Materials*, pp. 449–451, Aug. 1993.
[30] H. Kosaka, T. Shibata, H. Ishii, and T. Ohmi, "An excellent weight-updating-linearity synapse memory cell for self-learning neuron MOS neural networks," in *IEDM Tech. Dig.*, Dec. 1993, pp. 623–626.
[31] K. Kotani, T. Shibata, and T. Ohmi, "Neuron MOS binary-logic integrated circuits," in *Proc. Int. Conf. Advanced Microelectronic Devices and Processing*, Sendai, Mar. 3–5, 1994, pp. 609–614.
[32] M. Lenzlinger and E. Snow, "Fowler-Nordheim tunneling into thermally grown SiO_2 films," *J. Appl. Phys.*, vol. 40, no. 1, pp. 278–283, Jan. 1969.

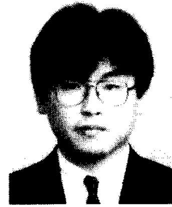
Authorized licensed use limited to: TOHOKU UNIVERSITY. Downloaded on March 01,2010 at 23:44:54 EST from IEEE Xplore. Restrictions apply.

[33] T. Shibata and T. Ohmi, "A self-learning neural-network LSI using neuron MOSFET's," in *Dig. Tech. Papers, 1992 Symp. VLSI Technol.*, Seattle, WA, June 1992, pp. 84–85.
[34] K. Hieda, M. Wada, T. Shibata, and H. Iizuka, "Optimum design of dual-control gate cell for high-density EEPROM's," *IEEE Trans. Electron Devices*, vol. ED-32, no. 9, pp. 1776–1780, 1985.
[35] J. L. McCreary and P. R. Gray, "All-MOS charge redistribution analog-to-digital conversion techniques—Part I," *IEEE J. Solid-State Circuits*, vol. SC-10, pp. 371–379, 1975.
[36] H. Ishii, T. Shibata, H. Kosaka, and T. Ohmi, "Hardware-backpropagation learning of neuron MOS neural networks," in *IEDM Tech. Dig.*, Dec. 1992, pp. 435–438.
[37] H. Ishii, T. Shibata, and T. Ohmi, "Hardware-oriented learning algorithm implemented on silicon using neuron MOS technology," in *Ext. Abst. 1994 Int. Conf. Solid State Devices and Materials*, Yokohama, Aug. 1994, pp. 382–384.
[38] S. Kondo, T. Shibata, and T. Ohmi, "Superior generalization capability of hardware-learning algorithm developed for self-learning neuron-MOS neural networks" *Japan J. Appl. Phys.*, vol. 34, pt. 1, no. 2B, pp. 114–117, 1995.
[39] D. E. Rumelhart *et al.*, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, Vol. I, D. E. Rumelhart and J. L. McClelland, Eds. Cambridge, MA: M.I.T., pp. 318–362.
[40] O. Fujita, Y. Amemiya, and A. Iwata, "Characteristics of floating gate device as analogue memory for neural networks," *Electron. Lett.*, vol. 27, no. 11, pp. 924–926, 1991.
[41] K. Kotani, T. Shibata, M. Imai, and T. Ohmi, "Clocked-Neuron-MOS logic circuits employing auto-threshold-adjustment," *ISSCC Dig. Tech. Papers*, Feb. 1995, FP 19.5, pp. 320–321.
[42] H. Kosaka, T. Shibata, H. Ishii, and T. Ohmi, "An excellent weight-updating-linearity EEPROM synapse memory cell for self-learning neuron-MOS neural networks," *IEEE Trans. Electron Devices*, vol. 42, pp. 135–143, 1995.

Tadashi Shibata (M'79) was born in Hyogo, Japan, on September 30, 1948. He received the B.S. degree in electronic engineering and the M.S. degree in material science from Osaka University, Osaka, Japan, and the Ph.D. degree from the University of Tokyo, Tokyo, Japan, in 1971, 1973, and 1984, respectively.

From 1974 to 1986, he was with Toshiba Corporation, where he was a Researcher on the R&D of device and processing technologies for ULSI's. He was engaged in the development of microprocessors, EEPROM's, and DRAM's, especially in the process integration and the research of advanced processing technologies for their fabrication. From 1984 to 1986, he was a Production Engineer with one of the most advanced manufacturing lines of Toshiba. During the period of 1978 to 1980, he was a Visiting Research Associate at Stanford Electronics Laboratories, Stanford University, Stanford, CA, where he studied laser beam processing of electronic materials including silicide, polysilicon, and superconducting materials. Since 1986, he has been Associate Professor with the Department of Electronic Engineering, Tohoku University, and has been engaged in the research and development of ultra-clean technologies. His main interest is in the area of low-temperature processing utilizing very-low-energy ion bombardment for the promotion of processes as well as in the development of the ultra-clean ion implantation technology to form defect-free ultra-shallow junctions by low-temperature annealing. Since the invention of a new functional device *Neuron MOS Transistor* (νMOS) in 1989, he has been intensively working on the exploration of new architecture electronic circuits using νMOS'.

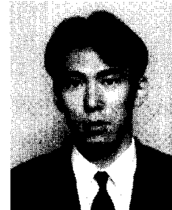
Dr. Shibata is a member of the Institute of Electronics, Information, and Communication Engineers of Japan, Japan Society of Applied Physics, and the IEEE Electron Device Society.



Hideo Kosaka was born in Saitama, Japan, on April 5, 1967. He received the B.S. and the M.S. degrees in electronic engineering from Tohoku University, Sendai, Japan, in 1992 and 1994, respectively.

When he was a graduate school student at Tohoku University, he studied the hardware implementation of neural networks using Neuron MOS transistors. In 1994, he joined the Research Center of Sony Corporation, Atsugi, Kanagawa, Japan, where he is currently working on the circuit design of microprocessors.

Mr. Kosaka is a member of the Institute of Electronics, Information, and Communication Engineers of Japan.



Hiroshi Ishii (M'88) was born in Tokyo, Japan on July 5, 1965. He received both the B.S. and the M.S. degrees in polymer chemistry from Kyoto University, Kyoto, Japan, in 1989 and 1991, respectively.

In 1991, he joined Nippon Motorola Ltd., Tokyo, Japan. From 1992 to 1993, he was a Visiting Researcher at the Electronic Engineering Department, Tohoku University, Sendai, Japan, where he studied the hardware implementation of neural networks using Neuron MOS transistors. He is currently working on the development of advanced ULSI process and device technologies at Tohoku Semiconductor, Sendai, Japan.

Tadahiro Ohmi (M'81) was born in Tokyo, Japan, on January 10, 1939. He received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Tokyo Institute of Technology, Tokyo, Japan, in 1961, 1963, and 1966, respectively.

Prior to 1972, he served as a Research Associate with the Department of Electronics at the Tokyo Institute of Technology, where he worked on Gunn diodes such as velocity overshoot phenomena, multivalley diffusion and frequency limitation of negative differential mobility due to an electron transfer in the multivalleys, high field transport in semiconductors such as unified theory of space-charge dynamics in negative differential mobility and block oscillators, and dynamics in injection layers. He is presently a Professor with the Department of Electronics, Faculty of Engineering, Tohoku University, where he is engaged in research on high-performance ULSI such as ultra high speed ULSI; Current Overshoot Transistor LSI, HBT LSI and SOI on metal substrate; base store image sensor (BASIS) and high speed flat panel display; and advanced semiconductor process technologies, i.e., ultra clean technologies such as high quality oxidation. Also, high quality metallization by low kinetic energy particle bombardment; very low temperature Si epitaxy having simultaneous doping capability by low kinetic energy particle bombardment; high-crystallinity film growth technologies from single crystal; grain size controlled polysilicon and amorphous due to low kinetic energy particle bombardment; in situ wafer surface cleaning technologies due to low kinetic energy particle bombardment; highly selective CVD; highly selective RIE; high quality ion implantation having low temperature annealing capability, etc., based on the new concept supported by newly developed ultra clean gas supply system; ultrahigh vacuum compatible reaction chamber with a self-cleaning function; ultraclean wafer surface technology; etc.

Dr. Ohmi has 350 original papers and 350 patent applications. He received the Ichimura Award in Industry-Meritorious Achievement Prize in 1979, Teshima Award in 1987, Inoue Harushige Award in 1989, the Ichimura Prize in Industry-Meritorious Achievement Prize in 1990, the IEEE Best Paper Award in 1990, and the Okochi Memorial Technology Prize in 1991. He serves as the president of the Ultra Clean Society. He is a member of the Institute of Electronics, Information, and Communication Engineers of Japan, the Institute of Electrical Engineers in Japan, the Japan Society of Applied Physics, and ECS.

Authorized licensed use limited to: TOHOKU UNIVERSITY. Downloaded on March 01,2010 at 23:44:54 EST from IEEE Xplore. Restrictions apply.