A Solid-State Neuron for Spiking Neural Network Implementation

Yajie Chen*, Steve Hall*, Liam McDaid†, Octavian Buiu*, and Peter Kelly†

Abstract—**This paper presents a compact analog neuron cell incorporating an array of charge-coupled synapses connected via a common output terminal. The novel silicon synapse is based on a two stage charge-coupled device where the weighting functionality can be integrated into the first stage. A presynaptic spike to the second gate allows the charge under the first gate to drift onto the floating diffusion output stage to produce a current, or voltage spike. Parallel defined synapses are each assigned to the left hand side of a current mirror gate where the right hand side feeds into a thresholding inverter. The decay of the membrane potential is mimicked by the charge leakage through a reverse-biased diode, whose model is verified by comparing the simulations and measured data. Spice simulation results show that the proposed neuron cell is capable of capturing the summing and thresholding dynamics of biological neurons.**

Keywords:* *neuromorphic circuits, silicon synapse, spiking neuron

1 Introduction

In recent years the fast progress in biological study has attracted a growing interest in mimicking the signal processing functions of biological neural systems, which can offer solutions to mathematically intractable problems through their ability to be trained for specific tasks. Experimental evidence has revealed that the traditional mean firing rate method could not describe brain activity since the reaction times are too short to allow temporal averaging for the calculation of the mean firing rate. This has given rise to the interest of alternative coding techniques which has led to the recent trend of Spiking Neural Networks (SNNs), an essential component in information processing by the brain [1]. On the other hand, the ITRS roadmap for Si indicates that alternative paradigms for building computational machines will be required within about 10 years and the massive parallelism of neural systems represents an attractive option. Therefore much significant joint research has been carried out to develop very large scale, highly parallel hardware implementation techniques of biological neurons [2]-[4]. However, most implementation approaches fail to match the dimensions of biological systems, since each neuron contains many synaptic inputs ($\sim 10^4$) and the physical space occupied by synapses will far exceed that occupied by the neuron cell. Although a single transistor can be engineered to implement multiplication and local learning [5], their characteristics are restrictive for spiking neurons based synapses where the time constants associated with the output transient of the loaded synapse as well as plasticity play a vital computational role within the neuron cell. Therefore, there is a pressing need for small, low power neural building blocks with operational characteristics that closely mimic realistic neuron cells.

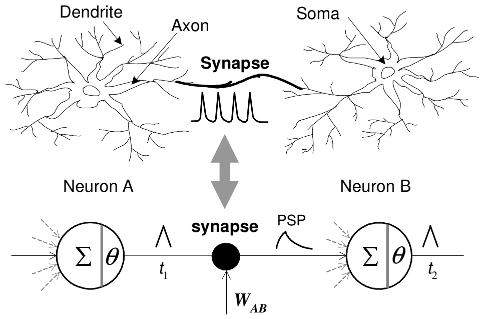

Figure 1: Illustration of a fragment of neural networks with synaptic junction. θ represents the action threshold.

Consider a fragment of a spiking neural network, so-called third generation (3G) artificial neural network [6], consisting of two point neurons with a synapse as shown in Fig. 1. At the synaptic junction a post-synaptic potential (PSP), whose magnitude is weighted according to W_{AB}, is generated in response to the spike emitted by neuron A. Note that the PSP is a transient with significantly different rise and fall time constants. This behaviour is caused by the loading effect associated with the post-synaptic (neuron B) membrane. At this membrane node, if the sum of the inputs from different dendrites surpasses a particular threshold then a spike is produced which propagates along the axon to other synapses. The firing properties are determined by a balance of synaptic excitation and inhibition. These biological characteristics provide the basis for this work.

*Department of Electrical Engineering & Electronics, University of Liverpool, Liverpool L69 3GJ, UK (Email: y.chen2@liverpool.ac.uk; s.hall@liv.ac.uk).

†School of Computing & Intelligent Systems, University of Ulster, Londonderry, BT48 7JL, UK.

In this paper, we present a biologically plausible neuron cell based on our recently developed charge-coupled synapse [7]. The silicon synapse is based on a two stage charge-coupled device (CCD) which is able to capture the dynamics of a biological synapse by using innate features of the semiconductor physics. The fundamental functionality of the biological neuron cell is implemented by current mirror summing, charge integration onto a thresholding inverter and subsequent slow leakage of charge via a reverse biased diode. Correspondence is made between the semiconductor relaxation processes and biologically relevant responses such as PSP and refractory period.

2 The Charge-Coupled Synapse

We now describe our charge-coupled synapse which is capable of mimicking the spiking dynamics of a biological synapse. Fig. 2 shows the n-channel, two-phase charge-coupled synapse, essentially consisting of two MOS capacitors in series. For the present study, we consider that a voltage, representing the weight, is placed on the first capacitor, but the intention in due course, is to include a floating gate upon which charge can be stored in a similar manner to that of non-volatile memory. The second MOS capacitor serves as the input node of the synapse which is triggered by the pre-synaptic signal. This device mimics the excitatory synapse. An inhibitory synapse could be implemented as a p-channel charge-coupled device. The output is formed by an n-type implant, on the right hand side of Fig. 2. The n-implant to the left can be used as a minority carrier injector, to speed up the non-equilibrium relaxation response of the device to achieve biological scale time regimes.

The synaptic weight is stored as the charge packet Q_{W} in the inversion layer of the first MOS capacitor. Note that the amount of charge increases linearly with the applied voltage on the first MOS capacitor which operates in the strong inversion state. Therefore the weight modification depends on the updating of the magnitude of the voltage V_{W}; weight storage will be considered in a subsequent publication. When a pulse is emitted by the pre-synaptic neuron, the silicon beneath the gate of second capacitor is driven into deep depletion and a deeper potential well is formed as shown in Fig. 3. Therefore Q_{W} drifts laterally under the gate and eventually onto the floating diffusion node (FDN). The charge Q_{W} can only be established through thermal generation of electron/hole pairs in the depletion region of capacitor 1, which is itself in deep depletion once the charge packet has gone. The lateral drift of charge onto the output node will result in a transient current spike at the output, limited by the charge density of the inversion layer, which diminishes with time. A refractory period is therefore implemented after the activation of the charge-coupled synapse because the recovery of the device from the deep depletion conditions takes the order of milliseconds when we employ lifetime quenching. Note that a pulse V_p applied to the left hand n+ region of Fig. 2, can be used to inject minority carrier electrons to facilitate and control this relaxation process in agreement with the biological frequency regime.

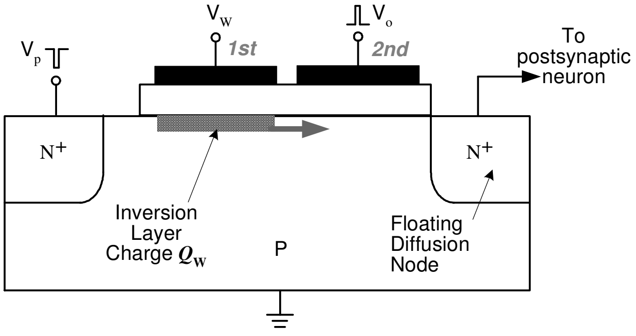

Figure 2: Schematic view of the charge-coupled synapse.

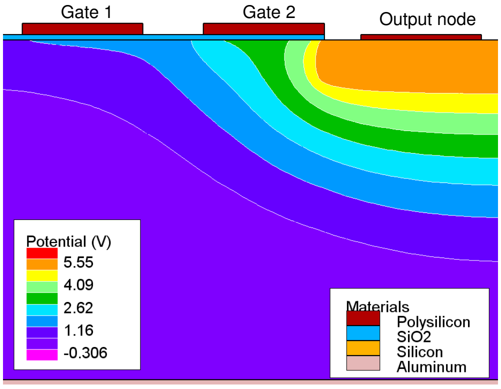

Figure 3: Potential profile of the activated synapse. The deeper potential under electrode 2 enables the charge transfer, causing a slope on the potential profile. Two-dimensional device simulation has been performed in Silvaco Atlas simulator. The substrate doping is $10^{15}\,\mathrm{cm}^{-3}$ and the N^{+} doping is $10^{19}\,\mathrm{cm}^{-3}$. The oxide thickness is $0.05\,\mu$m. The length of both electrodes is $1\,\mu$m with $0.5\,\mu$m spacing.

The arrival of the charge onto the FDN will induce a change in voltage given by:

$$\Delta V(t) = \frac{Q_{\mathrm{W}}}{C_{\mathrm{FN}}(t)} \tag{1}$$

where $Q_{\mathrm{W}}{=}\eta q N_{\mathrm{W}} L_1 W$ is the charge packet under the first electrode which has length L_1 and width W; η is the transfer efficiency; C_{FN} is the capacitance of FDN. Because of the coupling effect between the second electrode and FDN, there is a coupling voltage dropped onto the FDN. Thus the total transient voltage appearing on the FDN will exhibit the time-dependent properties, as shown in Fig. 4(a). The rise of the voltage strictly follows the increase of the presynaptic signal on electrode 2, and the fast rise time is dictated by the dielectric relaxation

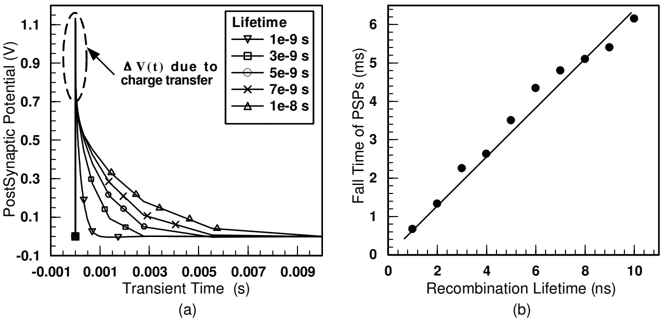

Figure 4: (a) The lifetime-dependent FDN potential mimicking the biological PSPs due to the activation of the synapse; (b) The decaying times of the PSPs for different recombination lifetimes.

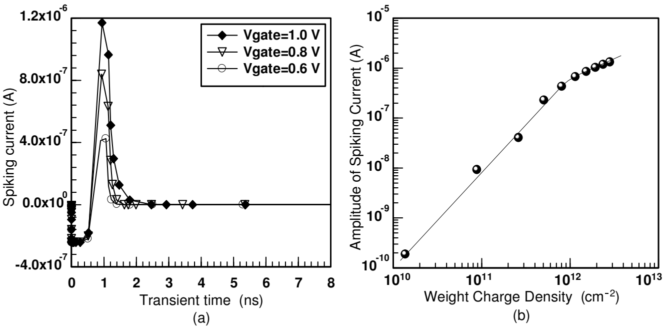

Figure 5: (a) Spiking current; (b) Amplitudes of spiking current as a function of weight voltage V_{W}. The oxide thickness is 25 nm and the electrode spacing is 0.25 μm.

time for majority carriers in the FDN and substrate. The falling edge of the spike is dictated by two processes: the arrival of the electron charge packet with a short time duration τ_1, and the subsequent lifetime-dependent relaxation of the voltage by leakage in the junction formed by the FDN and the substrate with time duration τ_2. As shown in Fig. 4(b), the latency of the FDN potential decreases with the decreasing lifetimes. The time τ_2 is therefore likely to be prohibitively long without lifetime quenching using techniques employed in power devices. PSP duration of milliseconds is easily achievable by these means. For the 1 ns lifetime, the latency is about 0.667 ms; whereas the latency increases to 6.15 ms for the lifetime of 10 ns. This linear relationship provides a possible engineering solution for the biologically plausible silicon synapse, which is applicable and compatible with different kinds of silicon neuron cell for building biomimic neural networks.

Generally, when the charge-coupled synapse is integrated into the neuron cell circuit described in the next section, the dc bias on the FDN will aid the collection of charge packet. In this case, the synaptic output is read as a current spike, as shown in Fig. 5(a). For a series of weight voltages from 0.6 V to 1.0 V, the variation of spike amplitude is approximately linear in agreement with conventional MOS physics. Fig. 5(b) shows the correlation between spike current amplitude and weight charge concentration. This relationship is used to implement synaptic plasticity in the proposed silicon synapse.

The charge-coupled synapse is able to perform the learning process by incorporating a floating gate on the first MOS capacitor instead of a poly-Si gate. It should be noted that two control lines are required to initiate updating of the weight: namely a feedback line from the post-synaptic neuron output, and the pulse from the presynaptic neuron which is also applied to capacitor 2. By this means, the learning can be implemented whereby an input pulse to a synapse that causes the neuron to fire, is 'rewarded' by an increased weight. A solution for such a floating gate device has been reported [8].

3 The Spiking Neuron

In ideal spiking neurons if the sum of the inputs, from different dendrites, surpasses a particular threshold then a spike is produced which propagates along the axon to other synapses. The integrate-and-fire (I&F) neuron model consists of a capacitor and a threshold device [1]. Fundamentally the cell membrane acts as a capacitor where the potential of the cell membrane can be modeled as the response of a capacitor to an injection of current. In response to a stimulant current the capacitor is charged, and when the potential reaches level, a spike is produced and the potential subsequently returns to the resting potential. The concept presented here however, represents a new paradigm to realize a spiking neuron, which has the potential to build more biologically plausible networks.

The circuit diagram of the proposed silicon neuron cell is shown in Fig. 6. M_1-M_2 constitutes a current mirror which is used to integrate the weighted current spikes from a number of n-type charge-coupled synapses. M_3-M_4 forms a CMOS inverter that thresholds the summed signals and generates an output hi-lo transition indicating that the neuron has fired. A 'leaky' diode D_1 enables charge leakage that effectively mimics the PSP decay of biological neurons. M_7 is controlled by the feedback of the neuron output of the second CMOS inverter M_5-M_6, and serves to reset the neuron cell after firing. The second CMOS inverter gives a lo-hi transition and is the actual output which is fed to subsequent synapses.

The charge-coupled synapses are directly connected to the drain of M_1 via a common output terminal. Note that the leakage of charge through the p-n junction formed by the output terminal and the substrate in the silicon synapse is controlled by the thermal generation lifetime and this process is orders of magnitude greater than the aforementioned charge transfer time. Therefore we can

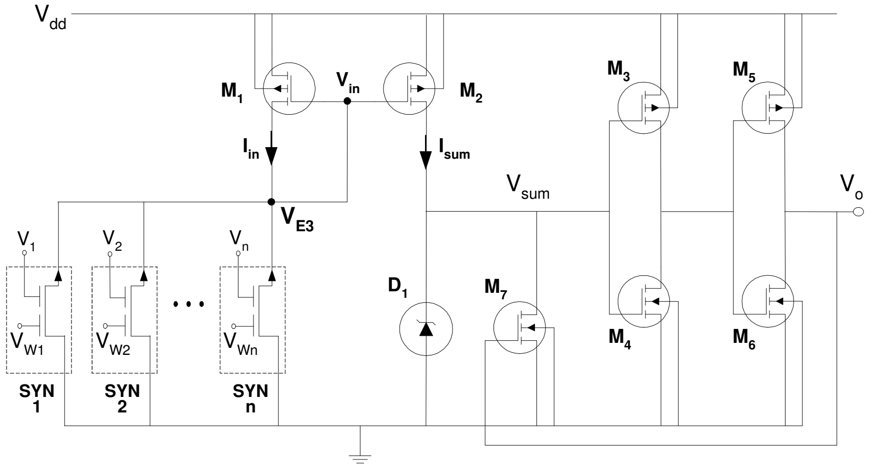

Figure 6: The analog neuron circuit with an array of n-type charge-coupled synapse; The integration and thresholding functions are implemented by the current mirror configuration and CMOS inverter respectively.

consider that no charge is lost during the summing operation.

3.1 Current Mirror Operation

All n synaptic output signals are integrated onto the drain of M_1. The transfer of weighted charge packets in n-connected silicon synapses results in a current I_{in}. Two p-channel devices which can be triggered by I_{in} are required in the current mirror configuration. Therefore the accumulated charge is transferred to the drain of M_1, the reference current I_{in} increases and is mirrored as I_{sum} in M_2. The current I_{sum} charges the gate voltage V_{sum} of the first CMOS inverter and D_1 is now under reverse bias.

Assume that two p-channel MOSFETs, M_1 and M_2, have the same oxide thickness and threshold voltage. The current for M_1 (assuming a long channel device) in saturation, whose source and substrate is connected to V_{dd}, is given by:

$$I_{in} = k_1 \frac{(V_{GS} - V_T)^2}{2} \tag{2}$$

where $k_1 = \mu C_{ox} \frac{W_1}{L_1}$ is the gain factor of M_1, and $V_{GS} = V_{in} - V_{dd}$.

The summed current from the synaptic array sets the V_{in} to:

$$V_{in} = V_T + V_{dd} - \sqrt{\frac{2I_{in}}{k_1}} \tag{3}$$

By using appropriate W_1/L_1 ratio, larger fan-in of the current mirror can be achieved, allowing the integration of a large number of synapses. For the device with equal W_1 and L_1, 20 nm oxide, and the power supply of 3 V, the approximated number of synapses in one-dimensional array can be 320. M_2 is then controlled by V_{in}, and the output current I_{sum} is obtained:

$$I_{sum} = k_2 \frac{(V_{in} - V_{dd} - V_T)^2}{2} \tag{4}$$

where $k_2 = \mu C_{ox} \frac{W_2}{L_2}$ is the gain factor of M_2.

Substitute (2) into (3), the current I_{sum} is expressed as a function of I_{in}:

$$I_{sum} = \alpha I_{in} \tag{5}$$

where $\alpha = \frac{W_2 L_1}{W_1 L_2}$ is the aspect ratio. This allows current signals to have a fan-out greater than one and each output can be scaled using appropriate W/L ratios. If M_1 and M_2 are matched in all respects, $I_{in} = I_{sum}$. After the spike emission period, V_{in} is driven back to V_{dd} causing the resetting of the common output terminal of the synaptic array.

3.2 Thresholding Operation

The CMOS configuration is employed in the circuit to perform the thresholding function which plays an important role in neuron computation. The summed current I_{sum} will charge the gates of M_3-M_4, causing the membrane voltage V_{sum} to increase. Consider the case where only one synapse is activated, and assuming a charing duration of τ_r. Therefore the membrane voltage can be expressed as:

$$V_{sum} = \frac{1}{C_{ON}} \int_0^{\tau_r} I_{in} dt \tag{6}$$

where $C_{ON} = C_D + C_p$ is the capacitance at the membrane node; C_D is the diode capacitance; C_p is the capacitance associated with the CMOS input gates.

The discharging of the gates of M_3-M_4 is due to the leakage through the reverse-biased diode D_1, as shown in Fig. 6. A 'leaky' diode with a reverse-current characteristic dominated by Zener tunneling is preferred and results in biological scale relaxation rates. A numerical solution gives us the approximation that $\tau_r << t(V_{sum})$.

The switching threshold V_{Th} for CMOS inverter is defined at the point where the input and output voltages are equal. Assume the voltage supply is high enough so that the transistors operate in saturation. The expression for V_{Th} is given by:

$$V_{Th} = \frac{(V_{T4} - \frac{V_{DSAT4}}{2}) + r(V_{dd} + V_{T3} + \frac{V_{DSAT3}}{2})}{1 + r} \qquad (7)$$

where V_{T3} and V_{T4} are the threshold voltage of pMOSFET and nMOSFET; V_{DSAT3} and V_{DSAT4} are the saturation drain voltages for M_3 and M_4 respectively; $r = \frac{k_3 V_{DSAT3}}{k_4 V_{DSAT4}}$ assuming equal oxide thickness; k_3 and k_4 are the gain factors.

When I_{in} increases sufficiently to make V_{sum} exceed the switching threshold of the CMOS inverter, M_3 is on and in saturation and M_4 operates in the linear region. The output voltage is then located at low level, which means the neuron has fired. Subsequently the output voltage V_o of the neuron will increase rapidly and turn on the transistor M_7. This discharging action resets the inputs of M_3-M_4, and the output V_o is constrained by the feedback to be a very short voltage spike.

3.3 Simulation Study

The silicon neuron circuit is simulated in Spice. The models are defined to match the devices fabricated in 0.5 μm process. The oxide thickness of the accompanying transistors is set to 20 nm which is used for charge-coupled synapse to provide enough fringing field. The threshold voltages of the p- and n-channel transistors are -0.74 V and 0.74 V respectively. For all the transistors, the same width and length are used (W=L=1.2 μm). Note that the size of the transistors can be varied to demonstrate the exact behavior of the realistic biological neurons. We represent the output of the charge-coupled synapses using current pluses of amplitude 1.2 μA for each activated synapse. The power supply V_{dd} is set to 3 V.

A low breakdown 'leaky' zener diode, whose reverse leakage is dominated by Zener tunneling, is employed to form a leakage path. This allows the discharge of the PSP which mimics the decaying of the membrane potential in biological neurons. Fig. 7 shows the I-V curves of the zener diode under non-zener condition. Good correlation between the modeled and measured characteristic is evident.

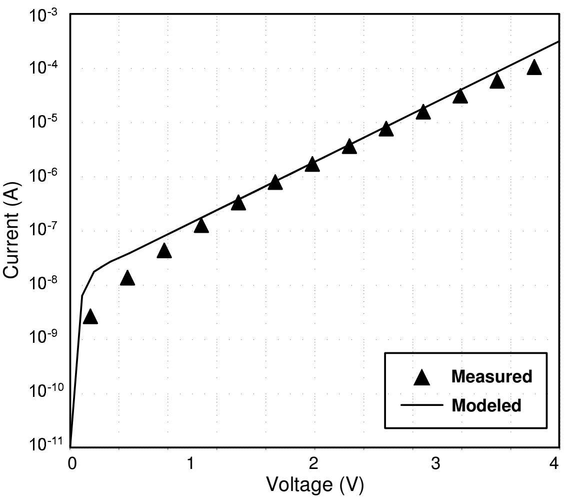

Figure 7: Measured and modeled I-V curves of the reverse-biased zener diode with breakdown voltage of 4.7 V.

In our experiment, 100 synapses were activated concurrently and the accumulated peak in current is 120 μA. The membrane node voltage V_{sum} at the gates of M_3-M_4 is shown in Fig. 8. The voltage, following the rise of the synaptic signal, has a fast rise time reaching a peak value 2.7 V which causes the first CMOS inverter to change state as shown in Fig. 10(b), demonstrating the firing of the biological neuron. Due to the leakage of D_1, the PSP shown has a much slower fall time of the order of milliseconds, exhibiting the characteristic shape observed in biological neurons.

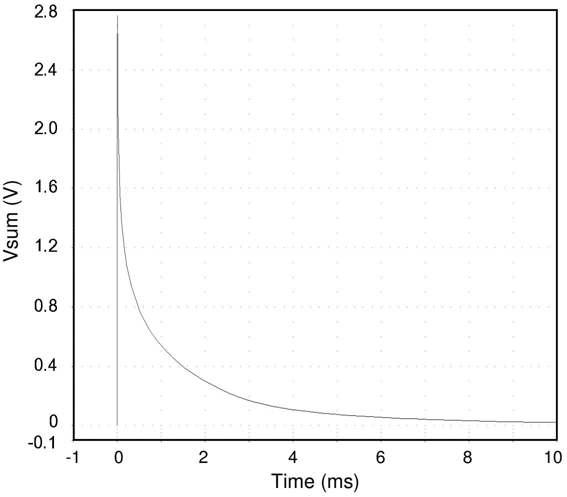

Figure 8: The membrane potential V_{sum} when the spikes are generated at the same time.

An estimate of the time required for V_{sum} to decay away is plotted in Fig. 9. The substrate doping level is 10^{15} cm^{-3}. From the shape of the plots obtained with the lifetimes of 1 μs and 0.1 μs, it can be concluded that for smaller spikes the decay time of V_{sum} is proportional to $\sqrt{V_{\text{sum}}}$, while for larger spikes the decay time can be expressed as a function of $\ln(V_{\text{sum}})$, which is more characteristic of biological neurons.

Fig. 10(a) shows the accumulation of successive PSP sig-

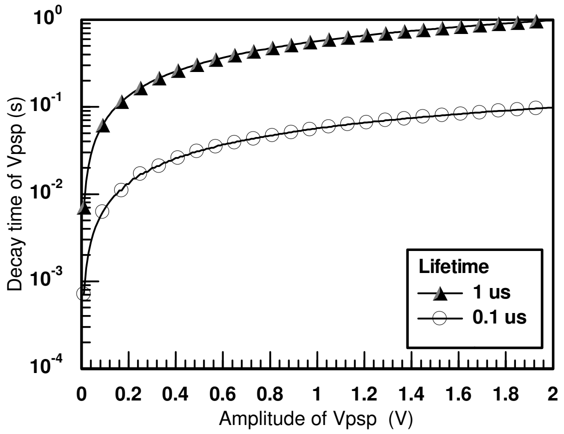

Figure 9: Total decay time of V_{sum} as a function of the amplitude of V_{sum} with generation lifetimes τ_g=1 μs and 0.1 μs.

nals when the synapses emit spikes with various time lags. In this simulation, there are 30 active synapses divided into 6 groups where each synapse is activated at the same time with an interval between each group. The associated accumulation of charge produces the V_{sum} shown in Fig. 10(a). Therefore, the threshold is reached and the CMOS inverter is triggered to send out a spike, as shown in Fig. 10(b). Because the membrane node discharging is of the order milliseconds, the accumulation behavior is limited by the interval between each group. If we assume 2 ms interval, the majority of charges are used to compensate the leakage rather than commit to neuron's firing. In addition, any spike arriving after the resetting of the membrane node will contribute to the subsequent firing events. It is worth noting that the resetting signal to the gate of M_7 should have a certain delay to ensure the neuron output V_o has a fan-out capable of driving multiple synapses on subsequent layers.

4 Discussion and Conclusion

Referring to the solid-state neuron cell, the resetting of the membrane node is achieved by the turn-on of M_7 into which the neuron output is fed. In the case that a large V_{sum} arrives at the membrane node, the neuron fires immediately causing the neuron to reset too early rather than allowing slower relaxation via the leakage of the diode D_1. Thus the resetting signal to the gate of M_7 should have a certain delay [9].

To illustrate the packing density afforded by the proposed synapse, consider a conservative area estimate of 5 μm×5 μm (=25 $\mu\mathrm{m}^2$) for a charge-coupled synapse. The authors are aware that this is only approximate as interneuron interconnect and associated circuitry would have to be accounted for in the calculation. Nevertheless, if we assume that the total chip area is 1 cm^2 and make a conservative guess that only 10% of the total surface area is occupied by synapses, then it is easy to compute that 400 thousand synapses for a single layer planar process is possible. The neuron cell circuit is designed with only 7 transistors which consumes much less area compared to the current estimate reported in the literature [4].

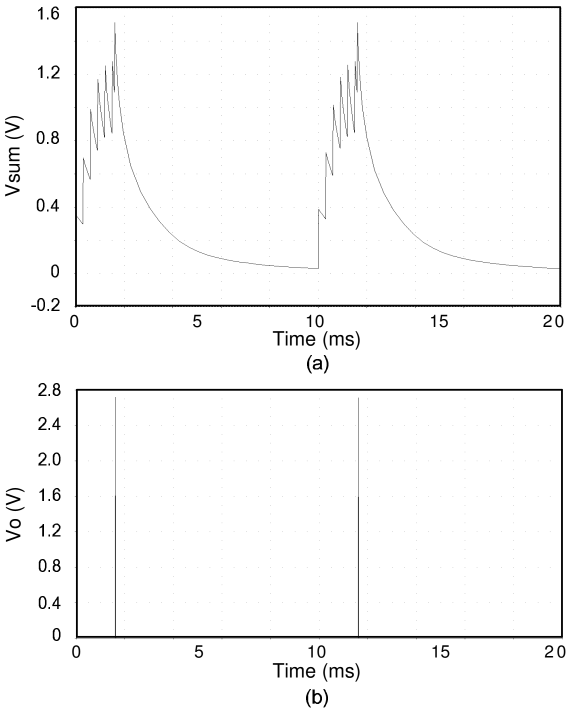

Figure 10: (a) The membrane potential V_{sum} when a series of spikes are generated with various time lags; (b) The output voltage of the second CMOS inverter M_5-M_6. The neuron fires when the membrane potential exceeds the threshold of CMOS inverter.

In conclusion, this paper presents device and circuit concepts for biological synapses and neurons in neural networks. By using the innate properties of semiconductors, a charge-coupled synapse capable of mimicking spiking dynamics and synaptic plasticity, serves as the compact component for implementing neural networks in silicon. The solid-state neuron presented, performs the integration of synaptic signals, captures the time dependency of the post-synaptic membrane decay, and fires when a certain threshold is reached. This silicon neuron cell together with an array of synapses will provide, for the first time, a core building block that is not only biologically plausible but has the potential to significantly advance the hardware implementation of spiking neural networks towards the biological-scale, using well proven and robust silicon technology.

References

[1] Gerstner, W., Kistler, W. M., *Spiking Neuron Models: Single Neurons, Populations, Plasticity*, Cam-

bridge University Press, 2002.

[2] Liu, S. C., Douglas, R., "Temporal Coding in a silicon network of integrate-and-fire neurons," *IEEE Transactions on Neural Networks*, V15, N5, pp. 1305-1314, 2004.

[3] Bofill-i-Petit, A., Murray, A. F., "Synchrony detection and amplification by silicon neurons with STDP synapses," *IEEE Transactions on Neural Networks*, V15, N5, pp. 1296-1304, 2004.

[4] Indiveri, G., Chicca, E., Douglas, R., "A VLSI array of low-power spiking neurons and bistable synapses with spike-timing dependent plasticity," *IEEE Transactions on Neural Networks*, V17, N1, pp. 211-221, 2006.

[5] Diorio, C., Hasler, P., Minch, B. A., Mead, C. A., "A single-transistor silicon synapse," *IEEE Transactions on Electron Devices*, V43, N11, pp. 1972-1980, Nov. 1996.

[6] Maass, W., "Networks of spiking neurons: The third generation of neural network models," *Neural Networks*, V10, N9, pp. 1659-1671, 1997.

[7] Chen, Y., Hall, S., McDaid, L., Buiu, O., Kelly, P., "On the design of a low power compact spiking neuron cell based on charge-coupled synapses," *IEEE International Joint Conference on Neural Networks*, Canada, pp. 1511-1517, 2006.

[8] Hieda, K., Wada, M., Shibata, T., Inoue, S., Momodomi, M., Iizuka, H., "Optimum design of dual-control gate cell for high-density EEPROM's," *IEEE Trans. Electron Devices*, VED-32, N9, pp. 1776–1780, Sep. 1985.

[9] Dowrick, T., McDaid, L. J., Hall, S., Buiu, O., Kelly, P., "An axon delay line using semiconductor junction leakage," unpublished.