Implementation of STDP in Neuromorphic Analog VLSI

Chul Kim
chk079@eng.ucsd.edu

Shangzhong Li
shl198@eng.ucsd.edu

Department of Bioengineering
University of California San Diego
La Jolla, CA 92093

Abstract

Spike-timing-dependent plasticity (STDP), an asymmetric form of Hebbian learning, shows how synaptic strength between neurons changes corresponding to time difference between pre- and post- spikes [1]. It is widely believed that synaptic plasticity can learn and store information of brain, so understanding STDP helps study of the process of learning in the brain. Moreover, hardware implementation of STDP is of great importance in developing brain- machine interfaces. In this paper, we simulate weight change respect to a fixed time difference in Matlab. Then we design circuits to investigate continuous-time STDP by showing weight changes between two neurons. The circuit, which includes integrate and fire (I & F) neuron module, synaptic trace module and weight tower module, is designed and simulated in the Cadence design environment. At last we compare the simulation results of circuits with Matlab simulation results.

1 Introduction

Herbbian learning rule is a biological theory which describes the mechanism of synaptic plasticity between neurons of brain during learning process. The synaptic plasticity efficacy can be enhanced while repeated pre-spiking neuron stimulates post neuron persistently. It can be described briefly as neurons fire together, wire together. However, the description is not quite exact. Technologies now make it possible to test the time difference between spiking of different neurons, thus induces the definition of STDP. With repeated pairs of pre- and postsynaptic spikes, if pre-spikes arrive a few milliseconds before post spikes, the synaptic strength is increased, which can cause long time potentiation (LTP). On the contrary, the strength is weakened if pre-neuron fires a few milliseconds after post-neuron, which can cause long term depression (LTD)[2].

Implementing STDP onto very-large-scale integration circuit makes it possible to model real electric behavior of neurons, which provides supports on designing brain-machine interfaces. In this paper, we present a circuit which combines I & F, spike tracing and weight module to display the weight changing respect to time. The simulation of STDP in Matlab is introduced in part 2, including LTP and LTD. In part 3, we describe all the circuits we design.

2 Simulation of STDP in Matlab

In [3], it has been proposed that the synaptic weight updates by each pre and postsynaptic firing events. Our model is based on this rule. When pre-neuron is fired, it undergoes the ions flowing

through the membrane. At the same time, it will release neuron transmitters to synapse cleft and bind to certain receptors of post neuron. Then the spike will leave a trace 'x'[4]:

$$\frac{dx(t)}{dt} = -\frac{x(t)}{\tau_+} \quad \text{where} \quad x = x + 1 \quad \text{when spike occurs} \qquad (1)$$

'X' is an abstract variable. Here 'x' can represent the number of transmitters that are bound to receptors. After each spike, the trace will decrease with a time constant τ_+. Similarly, we can also model the trace of post-neuron 'y' with a time constant τ_-:

$$\frac{dy(t)}{dt} = -\frac{y(t)}{\tau_-} \quad \text{where} \quad y = y + 1 \quad \text{when spike occurs} \qquad (2)$$

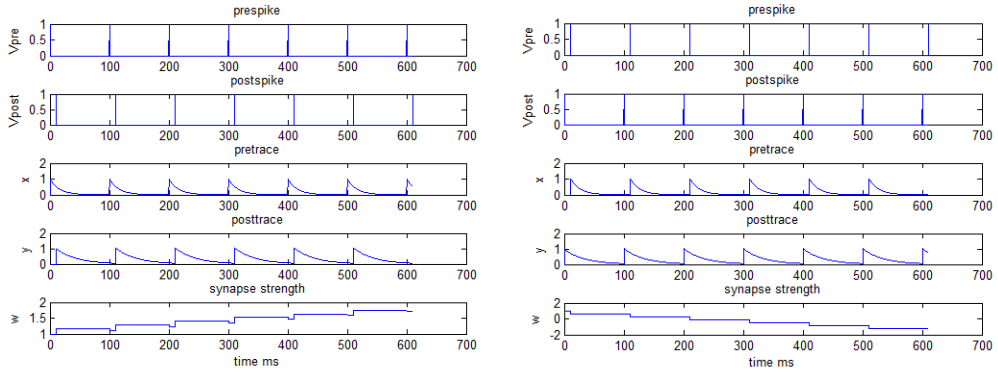
When pre-spike arrives, the trace of post-spike 'y' can be read out and decrease the weight:

$$\Delta w_d = -Ay(t) \qquad (3)$$

Conversely, 'x' can be read out based on the arrival of post-spike, leading an increase in weight:

$$\Delta w_d = \quad By(t) \qquad (4)$$

Simulation results are shown in figure1:



(a) (b)

Figure1. Simulation of STDP in Matlab. (a) LTP. Each pre-spike arrives 10 milliseconds ahead and results a little decrease in weight, while post-spike causes a big increase in weight. (b)LTD. Each pre-spike arrives 10 milliseconds later and causes a big decrease in weigh, while post-spike causes a small increase in weight. Here we set $\tau_+ = 16.8ms$, $\tau_- = 33.7ms$.

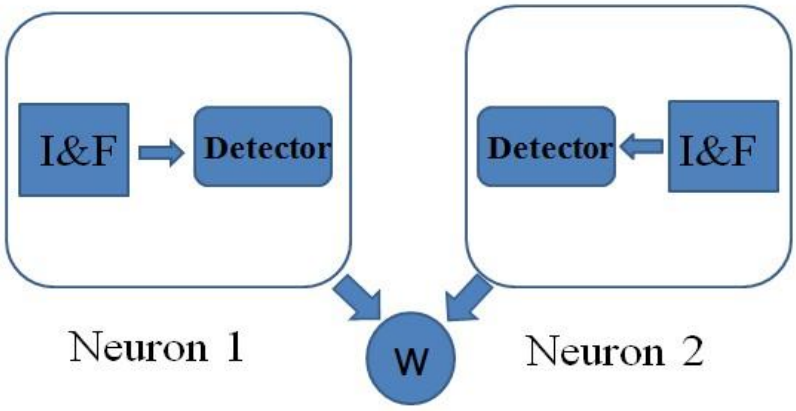


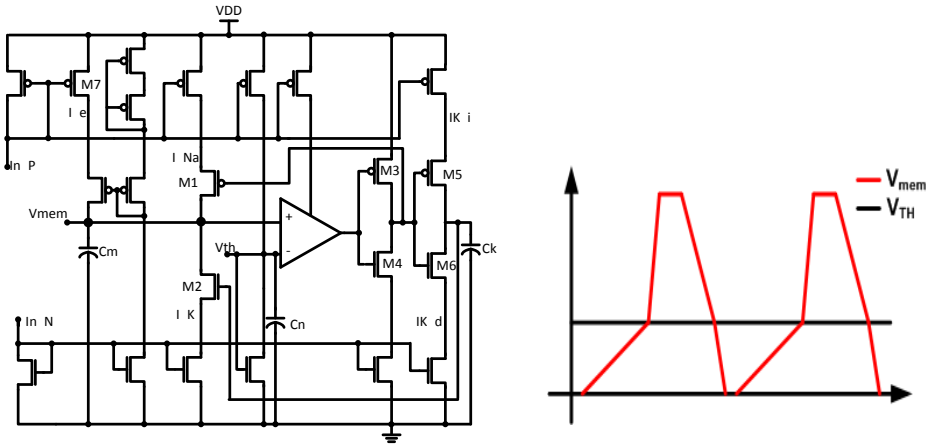
Figure2. Model of STDP. Each spike of the two neurons is generated by I&F module, the detector can display the traces, which can affect the weight.

3. Circuits

To achieve the implementation of STDP, the first thing is to come up with a model which can behave STDP. As STDP needs a pre-spike and post-spike, two neurons are needed at least. Our model is based on two neurons. Both neurons spike, and both of the spikes can trigger their own traces. Every spike can cause the weight change by using the previous trace of the other neuron. The scheme of the STDP model is described in figure2.

3.1 I&F module

In the biological way, the neuron spike cannot behave as a delta function which is shown in figure1. Actually, the neuron has a threshold for spiking, which means the stimulation must be larger than threshold to fire the neuron. In addition, after one spike, the neuron cannot be fired again until the voltage goes back to the resting potential. This period is call refractory period. All the neurons fire in biological way: the quick increase in potential is caused by sodium flowing into cell, and decrease caused by delay potassium flowing out.[5] I & F neurons is shown in figure3:



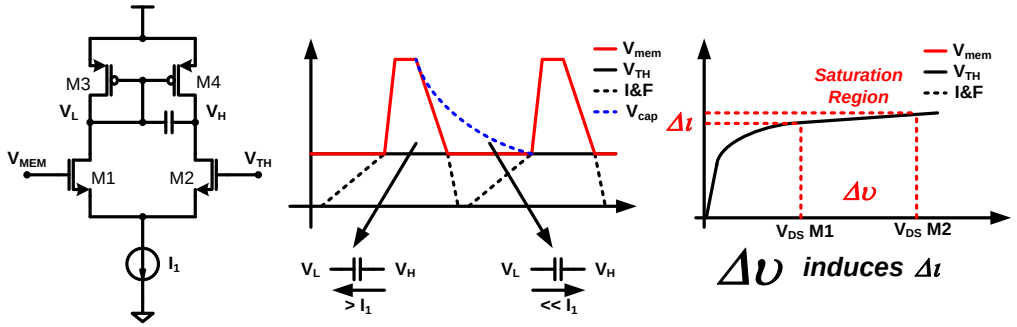
(a) (b)

Figure3. (a)The circuits of I & F module. (b) The theoretical simulation result.

In the circuit, In P and In N are the reference current inputs to provide current to this module. Cm models the biological neuron membrane. Vmem represents the voltage of neuron; Vth is the threshold voltage. Cn is utilized to reduce the noise. Ie acts as the excitatory input. If there is no current flowing into circuit through M7, nothing happens, which models the rest state of neuron. Once Ie exists, Cm will be charged, and Vmem will increase slowly. When Vmem is larger than Vth, the output of comparator is high. M3 and M4 combine together to serve as inverter. It is the same with M5 and M6. In this case, the output of the first inverter will be low, which can trigger M1 and enable I Na go through to charge Cm, causing Vmem increase rapidly. This is because I Na is bigger than Ie. Meanwhile, the output of second inverter will be high, which can charge Ck at a speed controlled by Iki. When the voltage of Ck is high enough, it will open M2 allowing Ik discharge Vmem. Ik is the biggest current in this module. As soon as Vmem is below Vth, the output of comparator is low, the output of first inverter is high, thus I Na, inhibited by M1, charges Vmem no more. Due to Ck, the output of second inverter moves to ground slowly. During that period, Vmem should be in ground. After discharging Ck enough, the current though M2 can be blocked and as a result, Vmem goes high slowly again using Ie current. Therefore, the refractory period can be defined by the Ie current and Cm.

3.2 Synaptic Trace

Each time neuron spikes, it produces a trace which follows an exponential decay. There are two kinds of trace: all-to-all trace and nearest trace. All-to-all means, when a pre-spike arrives at time t, the trace of pre-neuron will increase based on its value at the same time, and when a post-spike occurs, all of the pre-traces can be read out. The nearest trace confines the pre-trace increased from zero when a pre-spike arrives and the post-spike relates only the nearest pre-spike instead of all pre-spikes like our circuitThe circuit of trace module is shown in figure 4:



(a) (b) (c)

Figure4: (a) The circuit for trace. The capacitance is 0.5pF. (b) The voltage implemented by (a). (c) Characteristic curve of N channel field effect transistor.

Generally, sustaining voltage during 10ms is very hard using small capacitor.

$$\frac{1}{C}\int i\,dt = \frac{1}{C} i\Delta t = \Delta v \quad (5)$$

If we assume i in equation (5) is not changed, to sustain 200mV during 10msec with 0.5pF capacitor, i should be 10pA, which is almost impossible to make this small current intentionally using conventional way since small current like 10pA can be affected seriously by noise effect. Furthermore, even if it is possible to the small current like 10pA, this current cannot charge 0.5pF capacitor to 200mV within somewhat short period, 1ms. This means that we should make changeable current generator: when charging period, it current should be around a few nA, and when discharging period, around a few tens pA. To solve this problem, we proposed a new and simple circuit shown in figure 4, (a). In (a), we use the same Vth used in I&F module and set Vmem with a set of spikes whose trace is red line in (b). This Vmem is made from I&F module's spike. We use VH-VL to generate the synaptic trace like the blue line in (b). When a spike arrives at Vmem, M1 is open, allowing current flow; M2 is closed, inhibiting current goes through. At this time, the common source voltage between M1 and M2 can goes up and then, I1 can be increased due to intended non ideality. This increased current can flow from VDD to ground through M3 and M1, and this current can be copied by M4 due to current mirror structure. However, as the coppied current can't go through M2, it will flow through the capacitance Ct to charge it. As a result, the voltage through Capacitor, Ct, charges fast. The voltage of Ct doesn't change until Vmem starts to decrease. When discharging period (Vmem decreases and reaches Vth and stay until it goes up), discharging current can be very small due to the characteristic of MOSFET. If Vmem and Vth is exactly same each other, VL and VH want to have the same voltage. However, in this case, M1 and M2 have different V_{DS} as shown in (c) due to Ct charged before so thatVH is higher than VL, which leading the current through M2 is a little bit larger than M1. The current difference, Δi, is shown in (c). Owing to this characteristic, VH is discharged slowly, and ultimately reaches to the same voltage of as the blue line in (b). Therefore, we can only use small capacitor value, 0.5pF, for long time, 10ms.

3.3 W tower

With I & F spike and synaptic traces, the weight strength, here we call W tower, of synaptic plasticity can be generated by injecting repeated pairs of pre and post-spikes. When spike arrives, corresponded value of trace should be read out and be used to update the weight. From another aspect, the weight can only be updated when spike happens, at other time, it should maintain as a constant. The circuit of generating W tower is in figure5:

The left part and right part respectively represent post-spike trace and pre-spike trace and the middle part simulates the weight, which can operate only when either pre or post spike arrives. For example, when a post-spike occurs, M1 allows large current to be flown according to the strength of pre-spike trace, while M2 flows small portion of current due to the relative high value of VH.

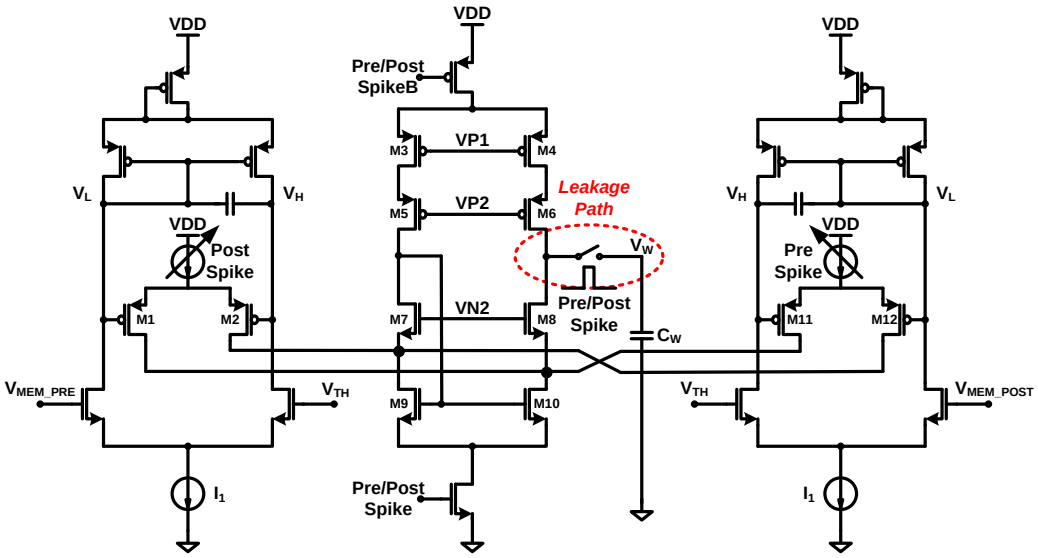


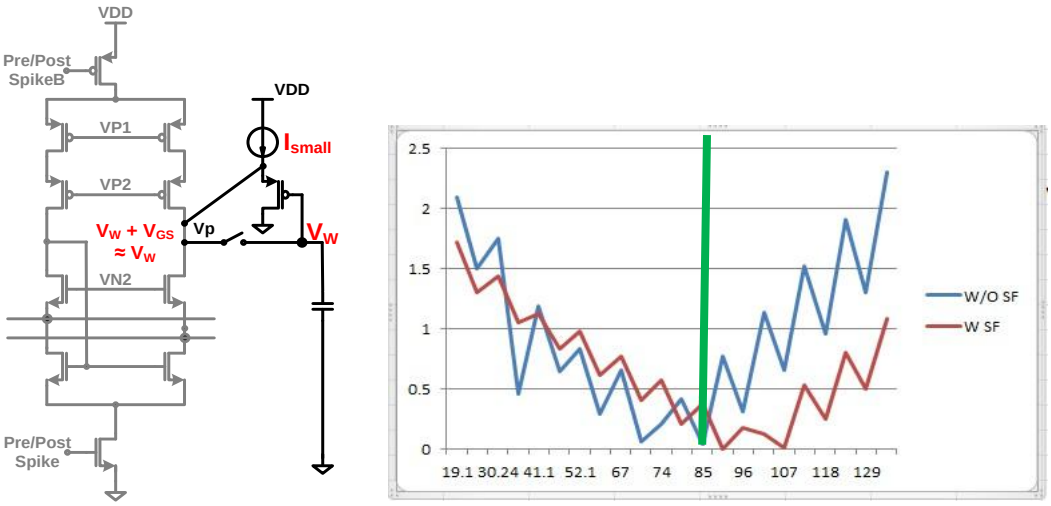
Figure5: W tower circuits with tracing circuits.

Normally, the amount of current from a pair of current source composed of 2 PMOS, M3 and M5 and M4 and M6, respectively, is the same. But, due to small current from M2, M9 and M10 can consume small current and at the same time, due to large current from M1, M10 get almost all current from M1 instead of M8. This leads to current path in Vw from the current source, M4 and M6. Therefore, the difference in current generated by two PMOS, M1 and M2, can go to C_W and Vw can be charged. This means that according to the time difference between pre and post spike, the voltage change in C_W can be different. Similarly, when a pre-spike arrives, the post-spike trace can be read out from the right part and then goes to the W tower. However, in this case, large current can be flown through M12, and as a result, M9 and M10 can absorb more current. This results in discharging C_W and decreasing V_W.

3.4 Leakage controller using source follower

The circuits introduced above can achieve modeling the STDP. However, it has an inherent problem: theoretically, weight can only be updated when spikes arrive and then be sustained, but in our circuit, as switches are not ideal, the leakage problem should occur and weight cannot maintain it as a constant long time.Generally, in order to deal with leakage problem, circuit designers should choose large capacitance. However, this is not a great way. Large capacitance means large area of silicon chip and finally, high cost. To avoid both leakage and large capacitance problem, in this project, we suggest a novel leakage control scheme, Leakage Controller using Source Follower (LCSF). LCSF is shown in figure6:

The main reason of leakage is non-ideality of switches which are usually made of NMOS and PMOS. Even though the gate voltages of two MOS do not allow switches to have channel for current path, if there is big voltage difference between Source and Drain nodes, this difference leads to strong E-field enough to overcome PN energy barrier. Furthermore, this leakage problem is getting serious as device shrinks to a few tens nm, such as 65nm CMOS. Our idea is originate from this main reason. If we reduce the difference in voltage between two nodes of switches, the leakage problem can be improved. To do so, we put additional circuits, source follower and a current source. The purpose of this circuit is, as mentioned earlier, making small voltage difference. Due to this circuit, Vp can be defined as V_W+V_{GS} by source follower instead of VDD or VSS when W tower is off. Since the current provided by current source is quite small, V_{GS} of M1 could be very small, which means the difference between Vw and Vp is almost negligible. The result of LCSF is shown in (b). When Vw is small, the advantage is not obvious, but when Vw is high, it shows the controller reduces leakage very well. This means, without LCSF, the voltage difference is very big when Vw is high. Total average of leakage slope changes from 0.9mV/ms to 0.6mV/ms.



(a) (b)

Figure6 (a) Leakage controller by using source follower: adding a PFET between V_W and VP.(b) The comparison of leakage between circuits with controller and without controller.

3.5 Timing circuit

When a spike of one neuron arrives, the value of detector of another neuron can be read out to update weight. Theoretically, only the detector value at the moment of spike should be used for updating weight. However, the spike will keep for some time, and during that time the value of detector is decaying, so it is good way to make the update time as little as possible. Another advantage for doing this is to reduce the consumption of the circuits since updating circuits only operates during update period. There is one more point for us to keep in mind. Generally, almost all circuits need time to be stabilized shortly after it is turned on. In order to avoid updating unstablized voltage, we should give time to stabilize to W tower. As a result, we need two pulse generators that make short pulse just after one neuron spikes. One is used for W tower and the other is for just switches.

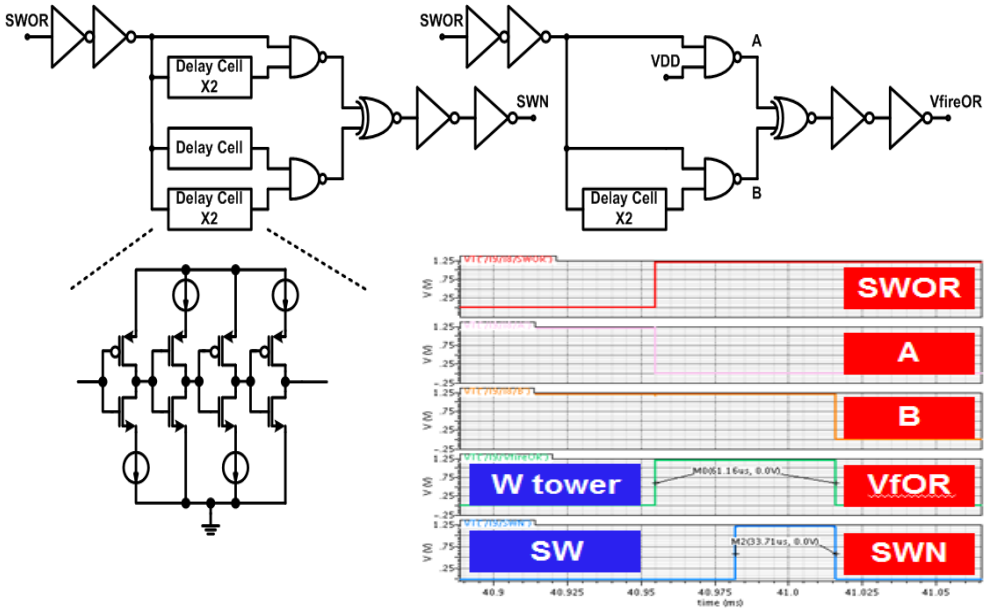
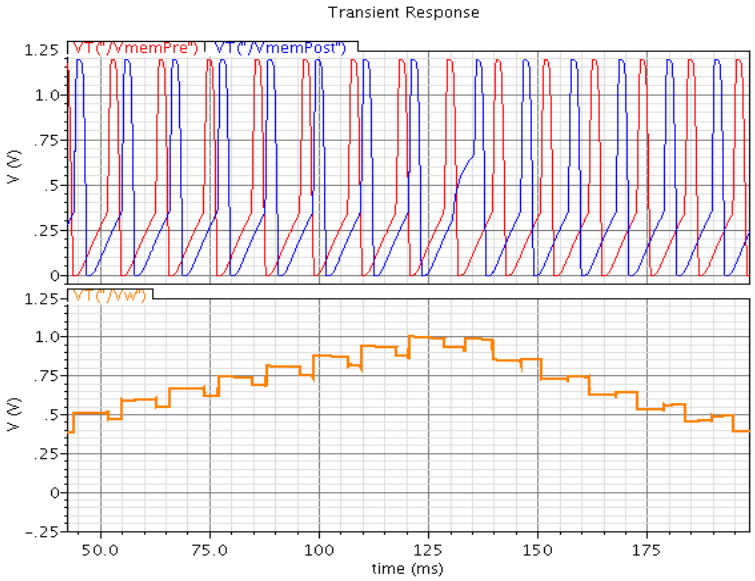


Figure7: (a) timing circuit. (b)simulation resut

We used the circuit shown in figure 7 to implement this function. The SWOR has information about spike. This signal goes up when one of neurons spikes. When SWOR is high, node A can be changed to low while B node is still high due to delay circuit. XOR gate make output voltage high when two input voltage, A and B, are different. So, VfireOR can be high after spikes. However, after delay, B can be low and output of XOR can be low due to the same input voltage. That is, the amount of delay can define the pulse width. Delay cell is composed of 4 inverter chain using small current source. This inverters transfer high voltage slowly, around 30us, whereas transfer low voltage almost instantly. The VFOR corresponds to the shortened spike, in other words, the w tower can only operate during the spike time of VFOR, 60us. After 30us , the SWN spikes which indicates the period of updating the weight.

4. Simulation result and conclusion



Capacitor	I&F	2.2p
	Det	0.5p
	W	0.9p
VDD	1.2V	
Power	0.94uW	
Process	65nm CMOS	

Figure8 (a) simulation result (b) parameters in our circuits

The upper figure shows the generation of spike by using I&F module. The integrated period, spiking period and refractory period can be identified clearly. The red line represents pre-spikes; blue line represents post-spikes. Before 125ms, we set pre-spikes arrive before post-spike, which causes LTP. The weight increases just as simulated in the Matlab. After 125ms, we set post-spikes arrive first, which causes LTD. Similarly, the weight decreases as simulated in the Matlab.

In the figure, it can be noticed that during the period without spike, the weight almost maintains in a certain value. This is benefit from using our leakage controller using source follower module which leads to a very little change in V_W during a quite large period of time. Furthermore, we can see that weight is updated almost at only one moment, this is because of our timing module, which shortened operation time of W tower and enables weight updating in 30us during which the value of trace can be treated as a constant.

Combining all of these advantages, our circuit can simulate the STDP very well. In figure 8 (b), we summarize important parameters used in this circuit. Owing to novel proposed circuits, we are able to use small capacitance compared to I&F circuit. For power consumption, we choose 1.2V for VDD and 65nm CMOS as a device. As a result, our circuits only consume under 1uW. However, if we use high VDD-3.0V- and old process-0.35um CMOS-like other chips for neuromorphic applications, there is possibility to make large voltage difference between switches enough to result in serious leakage problem. Then, our proposed scheme, LCSF, will more be highlighted due to its powerful effect for reducing leakage.

References

[1] Jesper Sjöström and Wulfram Gerstner (2010), Scholarpedia, 5(2):1362.

[2] Dayan P, Abbott L (2002) Theoretical neuroscience: computational and mathematical modeling of neural systems. Boston: MIT Press.

[3] Kistler WM, van Hemmen JL (2000) Modeling synaptic plasticity in con- Pfister and Gerstner • Triplets in STDP Models J. Neurosci., September 20, 2006 • 26(38):9673–9682 • 9681 junction with the timing of pre- and postsynaptic potentials. Neural Comput 12:385– 405.

[4] Pfister, J.-P. & Gerstner, W. Triplets of spikes in a model of spike timing–dependent plasticity.*J. Neurosci.* 26, 9673–9682 (2006).

[5] G. Indiveri, B. Linares-Barranco, T. J. Hamilton, A. van Schaik, R. Etienne-Cummings, T. Delbrück, S.-C. Liu, P. Dudek, P. Häfliger, S. Renaud, J. Schemmel, G. Cauwenberghs, J. Arthur, K. Hynna, F. Folowosele, S. Saighi, T. Serrano-Gotarredona, J. Wijekoon, Y. Wang and K. Boahen Neuromorphic silicon neuron circuits *Front. Neuromorph. Eng., vol. 5, pp. 1-23, 2011*