A Compact Spike-Timing-Dependent-Plasticity Circuit for Floating Gate Weight Implementation

A. Smith, L. McDaid and S. Hall

Abstract—**Spike timing dependent plasticity (STDP) forms the basis of learning within neural networks. STDP allows for the modification of synaptic weights based upon the relative timing of pre- and post- synaptic spikes. A compact circuit is presented which can implement STDP, including the critical plasticity window, to determine synaptic modification. A physical model to predict the time window for plasticity to occur is formulated and the effects of process variations on the window is analysed. The STDP circuit is implemented using two dedicated circuit blocks, one for potentiation and one for depression where each block consists of 4 transistors and a polysilicon capacitor. SpectreS simulations of the back-annotated layout of the circuit and experimental results indicate that STDP with biologically plausible critical timing windows over the range 10µs to 100ms can be implemented. Also a floating gate weight storage capability, with drive circuits, is presented and a detailed analysis correlating weights changes with charging time is given.**

I. Introduction

Significant research over the last two decades has been undertaken on studying biological neural networks. Specifically this research has focused on how neural networks learn and adapt to their ever changing environment together with the translation of this into biologically inspired hardware neural networks [1-2]. A neural network (NN) consists of interconnecting neurons, with each neuron connecting to another via a synapse. Within the human brain there are in excess of 10^{11} neurons, with each one having up to 10^3 synaptic connections [3].

In a NN, the effect that one neuron has upon another will vary depending upon input stimuli and synaptic weight. The synapse is responsible for adaption and learning within a NN [4], through long term potentiation (LTP) or long term depression (LTD), depending on the temporal ordering of the pre- and post-synaptic spikes. Additionally weight modification can also be a short term potentiation (STP) or a short term depression (STD).

Hebb's theory [5] describes how the synaptic weight is allowed to change based upon the inputs and outputs of each neuron within the NN. A further development of the Hebbian learning concept was the introduction of spike timing dependent plasticity (STDP) in 1983 [6]. STDP is concerned with increasing or decreasing the weight of a synapse based upon the relative timings of pre- and post-synaptic spikes. In biology two STDP functions are commonly reported and referred to as symmetric and asymmetric [4, 6-12]. In this paper we focus on asymmetric STDP as this type of plasticity is known to occur more frequently in biological NN, [4, 7, 11-12]. It is also worth noting that the exponential functions commonly depicted, are not a pre-requisite for STDP but rather a mathematical convenience. What is important however is the relative timings between pre and postsynaptic spikes as this temporal ordering dictates whether potentiation or depression occurs [46, 47]. In asymmetric STDP, weight potentiation (a pre-post spiking event) occurs if a pre-synaptic spike precedes the post-synaptic spike and this leads to LTP; Δt_s is positive. Likewise, the weight is decreased if a post-synaptic spike occurs prior to a pre-synaptic spike, giving rise to LTD (a post-pre spiking event, Δt_s is negative). The critical timing window [7, 14-18] typically occurs over the range 10-100msec and outside of this window, no potentiation or depression will occur [7, 14-20]. The critical timing window is implemented in this work and is programmable.

It has been shown that STDP can be implemented in hardware, and while the majority of these circuits are biologically plausible, their footprints are large [21-30] requiring up to and, in some cases, exceeding thirty MOSFETs. Other solutions require dedicated microprocessors. A key requirement of hardware neural networks (HNN) is that they are scalable and therefore the designs for neurons, synapses and synaptic modification circuits must be compact, low-powered, while at the same time maintain biological plausibility.

It is proposed here that an STDP circuit with critical time window can be implemented using two dedicated circuit blocks each consisting of 4 MOS transistors, and a polysilicon capacitor. The paper is organized as follows; in section II an overview of theoretical operation of the compact STDP circuit is presented. Section III presents experimental and simulation results undertaken in AMS 0.35µm CMOS process and SpectreS in the Cadence environment respectively. All simulations are conducted on back-annotated layouts, thus incorporating all parasitic elements. A discussion of results relating to the circuit properties is presented in section IV and conclusions drawn in section V.

II. Circuit Operation

This section provides an overview of the operation of the proposed STDP weight potentiation and depression circuits. Also a model for the critical timing window is given together with its dependency on process variations.

II.A WP and WD Circuits

The WP circuit is presented in Fig. 1(a). The circuit will cause an increase of the synaptic weight by increasing the amount of negative charge stored on the floating gate (FG) of a non-volatile memory device. This device is represented by its equivalent capacitance C_{FG}. The weight increase occurs during a pre-post spiking event. The WD circuit is identical to that of the WP block except that the pre and post spike input terminals are swapped. The WD circuit decreases the synaptic weight by removing charge on the FG during a post-pre spiking event.

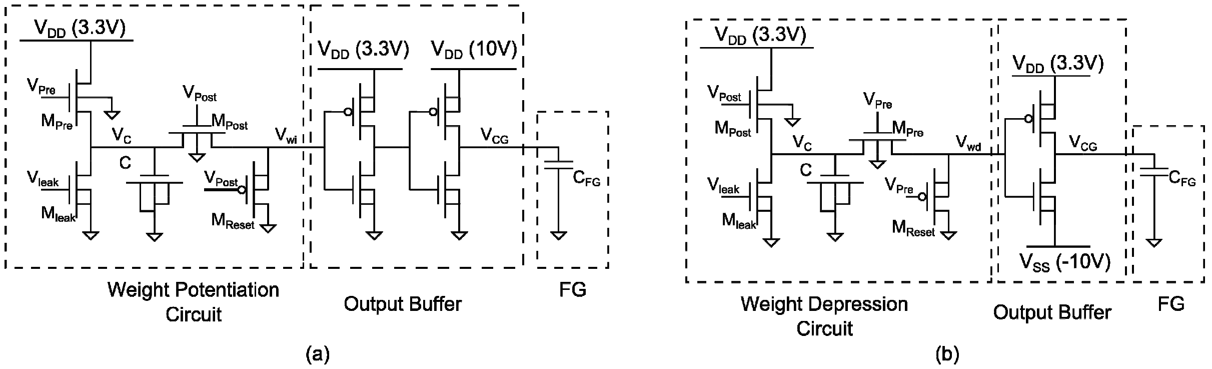


Fig. 1 (a) WP and (b) WD circuit block with FG device and driver buffer circuit. Voltages indicated are relative to ground.

The WP and WD circuits each consist of 3 NMOSTs, M_{Pre}, M_{Post} and M_{leak}, a PMOST, M_{reset} and a MOS capacitor, C. Transistor M_{reset} is used to ensure that, V_{wi} and V_{wd} are pulled low in the absence of V_{Post} and V_{Pre} respectively. When V_{post} and V_{Pre} are high, M_{reset} is off and will not significantly affect V_{wi} or Vwd. The operation of the WP circuit is now outlined. The initial conditions when no pre- or post- synaptic spikes occur are that $V_{wi,}$ V_{pre} and V_{post} are low, node V_C is pulled low by M_{leak} and C is discharged.

Consider a pre-post spiking event where a pre-synaptic spike (V_{Pre}), increases V_C to its maximum value (= 3.3V-V_{TMpre}): V_{TMpre} is the threshold voltage of M_{pre}. When the pre-synaptic pulse ends, C starts to discharge via M_{leak}, and V_C decreases at a rate determined by voltage V_{leak}. Voltage V_{leak} thus controls the timing window in which a post-synaptic spike must occur in order to cause the synaptic weight to be increased. When the post-synaptic spike (V_{Post}) occurs, the nodes with voltages V_C and V_{wi}, are connected and V_{wi} is pulled up to V_C -$V_{TMpost}(V_{wi})$; $V_{TMpost}(V_{wi})$ is the threshold voltage associated with M_{post}. The synaptic weight will be increased, while V_{wi} is greater than the trigger voltage of the output buffer.

The WP output buffer is constructed using two CMOS inverters with 3.3V and 10V V_{DD} rails, as shown in Fig. 1(a). The MOSFETs are sized so as to produce the following operation; if V_{wi} is greater than the trigger voltage of the first CMOS inverter then the output from the second inverter, $V_{CG,}$ will be pulled up to 10V. If V_{wi} is below the trigger voltage of the first CMOS inverter, then the output from the second inverter is held at ground. The pulse-width, τ_{cg}, and magnitude of V_{CG} determines how much charge is injected and stored on the FG. As $\Delta t_s \rightarrow \Delta t_{s\ min}$, $\tau_{cg} \rightarrow$ max τ_{cg}. Similarly as $\Delta t_s \rightarrow \Delta t_{s\ man}$, $\tau_{cg} \rightarrow$ min τ_{cg}. Finally for a post-pre spiking event no update of the synaptic weight occurs since V_C and V_{wi} are low, regardless of when the presynaptic occurs.

The operation of the WD block is similar to that of the WP block, with post-pre spiking causing a decrease in synaptic weight. The WD output buffer is constructed using a single CMOS inverter with 3.3V and -10V supply rails, as shown in Fig. 1(b). The inverter MOSFETs are sized so as to produce the following operation; when V_{wd}, is greater than the threshold voltage, the output of the buffer is pulled down to -10V. If V_{wd} is less than the threshold voltage of the inverter, then the output is 0V. For the case of pre-post spiking, the pre-synaptic spike causes V_C and V_{wd} to be pulled low and there is no update of the synaptic weight. It should be noted that if $\Delta t_s = 0$ (a pre- and post-synaptic spike occurring at the same time) then $\Delta w = 0$ because both the WP and WD circuits will be 'on' during this event causing node V_{CG} (Fig.1) to be set at 0V. This is consistent with biophysical experiments where it has been reported [50, 51] that synaptic communication between pre- and post-synaptic neurons is inherently delayed by axons or dendrite latencies and thus the actual strongest and weakest synapse efficacy does not occur at the absolute temporal difference ($\Delta t_s = 0$).

II.B Critical Timing Window

The critical timing window (CTW) is crucial in biology because it determines the time window over which synaptic modification can occur and is typically 20-25ms for potentiation and depression [7, 9]. However, in hardware the computational speed is greatly accelerated, with average spike train frequencies in the MHz range. We therefore implement an equivalent timing window of 20-25µs in this work although, as will be shown, the window can be programmed to accommodate a wide temporal range. We define here, the critical timing window, t_{cw}, as the time it takes for V_C to fall from 90% to 10% of its initial value for both the WP and WD blocks. The rate at which the sub-threshold current reduces V_C is set by V_{leak} and the aspect ratio of M_{leak}, S_{Mleak}. The sub-threshold current, I_{leak} is constant for $V_{DS} = V_C > 3kT/q$;

$$I_{leak} = \mu_{eff} C_o\, S_{Mleak}(m-1)\left[\frac{kT}{q}\right]^2 exp\left[\frac{q(V_{leak}-V_t)}{mkT}\right] \tag{1}$$

where V_t is the threshold voltage of M_{leak}, q is the charge of an electron, k is the Boltzmann constant and T is absolute temperature. The sub-threshold slope parameter, $m = 1 + C_d/C_o$ with C_d being the depletion layer capacitance, C_o is the capacitance of the oxide per unit area and μ_{eff} is the effective channel mobility. The dynamic operation of the capacitor charging is governed by $dt = -\frac{C}{I_{leak}}dV$, with I_{leak} given by Eqn.(1). Performing the integration with voltage limits, $0.9V_M$ and $0.1V_M$ gives equation (2) which can be used to determine the critical timing window, t_{cw}: V_M (= 3.3V-V_{TMpos}) is the maximum value of V_C. The window can be adjusted using V_{leak} according to:

$$t_{cw} = 0.8\frac{CV_M}{I_{leak}} \tag{2}$$

Substituting equation (1) into (2) and rearranging allows a value for V_{leak} to be calculated for the required t_{cw}. In this study, t_{cw} is chosen to be 20µs, giving V_{leak} =410mV.

The important effects of process variation upon the critical timing window are now considered. Process variation can affect most parameters of the MOSFET and these can conveniently be represented by the transconductance factor (β) and threshold voltage, V_t [31-43]. Subthreshold MOSFETs are particularly sensitive to process variation because of the exponential relationship between drain current and gate voltage (equation 1). The threshold voltage is also strongly related to several device parameters which are prone to variation during the fabrication process.

For M_{leak} operating in subthreshold, only V_t is considered, [35, 38, 43-44] as this incorporates variations in both off-current and subthreshold slope, as shown in equation (3), for an n-channel device, where N_a is the acceptor doping concentration, t_{ox} the oxide thickness, ϕ_F the Fermi potential, Φ_{MS} the work function difference, Q_t the trapped oxide charge density, C_o the oxide capacitance and ε_0, ε_s, ε_{ox} are the permittivity of free space, relative permittivity of silicon and silicon dioxide respectively.

$$V_{t0} = t_{ox}\frac{\varepsilon_s}{\varepsilon_{ox}}\sqrt{\frac{2qN_a(2\phi_F)}{\varepsilon_0\varepsilon_s}} + 2\phi_F + \Phi_{MS} + \frac{Q_t}{C_o} \tag{3}$$

The variation in $V_t = V_{t0} \pm \Delta V_t$where V_{t0} is the nominal threshold voltage for the AMS process, V_{t0} = 0.48, and $\pm\Delta V_t$ is the change in V_t due to process variations. For the AMS process ΔV_t = ±17.5mV. A simple model for the effect of process variation on t_{cw}, can therefore be written as:

$$\Delta t_{cw} = \frac{0.8V_M C}{I_0 exp\left[\frac{q}{mkT}(V_{leak}-[V_{t0}\pm\Delta V_t])\right]} \tag{4}$$

Monte Carlo analysis was undertaken in Cadence to assess the effects of inter-die/die-to-die process variation on the critical timing window and results are presented in Fig. 4. The results of Fig. 4, compare the Monte-Carlo simulations with equation (4), and good agreement is apparent with ΔV_t = ±17.5mV. The results also show a considerable change in the critical timing window, t_{cw}, from the ideal value of 20µs, due to process variation for V_{leak} = 410mV. For ΔV_t = +17.5mV, t_{cw} = 30.86µs, and for ΔV_t = -17.5mV, t_{cw} = 12.21µs.

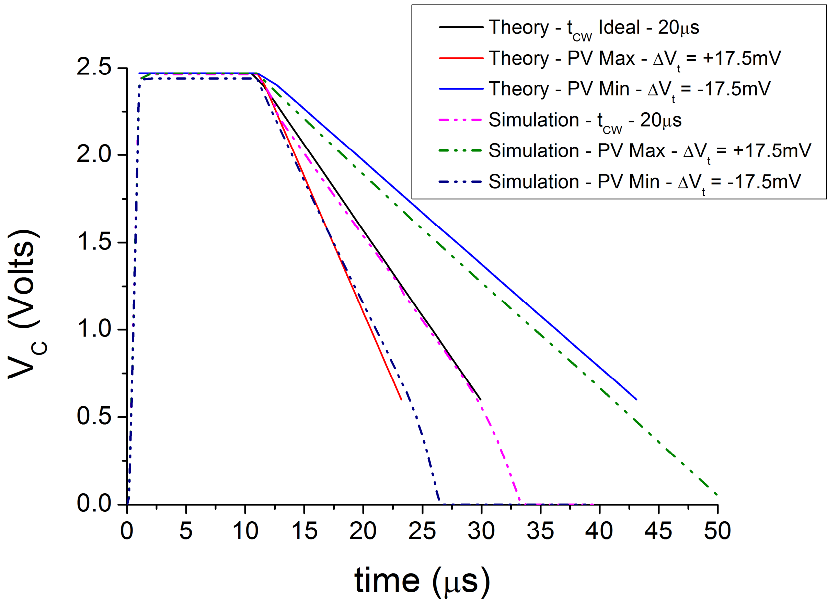


Fig. 2 t_{cw} variation (max, min and ideal) for V_{leak} = 400mV.

The effects of process variation on t_{cw} is presented later where it will be shown (Fig. 19) that this variation can be offset by adjusting the learning duration.

III. Results And Discussion

Simulation and experimental results for the WP block under post-pre spiking conditions are presented in section III.A. Simulated results for the WD block under post-pre spiking conditions are presented in section III.B In both sections III.A and III.B, V_{leak} is set to 410mV, C is 100fF (4.7µm x 4.7µm) and S_{Mleak} = 1 giving t_{cw} = 20µs from equation (5). Additional parameters for the circuit are; W_{Mpre} = L_{Mpre} = 0.5µm, W_{Mreset} = L_{Mreset} = 0.5µm, W_{Mpost} = 0.4µm L_{Mpost} = 0.35µm.

III.A WP Results

Fig. 3(a) presents simulation and measured results of a post-pre spiking event, where the pre-synaptic spike occurs 5µs after the end of the post-synaptic spike, Δt_s = 5µs. In this case no weight update occurs. This is because C is initially discharged with V_C = 0V due to the occurrence of the post spike before the pre spike. Results are now presented in Fig. 3(b), Fig. 4 and Table 1, for a series of pre-post spiking events where the time difference, Δt_s, between pre- and post- synaptic spike is increased from 1µs to 15µs. Fig. 3(b) indicates that V_{pre} causes C to be charged to voltage V_C =V_M, and then discharges to give t_{cw} = 20µs. Voltage V_{wi} tracks V_C after V_{post} occurs, triggering a weight update. It should be noted that V_{wi} is only pulled down to about V_t. For Δt_s =1µs, the maximum weight update occurs, $\Delta w = \Delta w_{max}$. This occurs as V_{wi} is above the trigger voltage of the output buffer, while V_{post} is still high. Thus V_{CG} is at its maximum pulse width, τ_{cg} = 10.91µs (simulation) and has a measured value of τ_{cg} = 10.75µs. In both cases V_{CG} has a magnitude of 10V. Fig. 5(b) shows that the measured value for V_C shows good agreement with the simulation results.

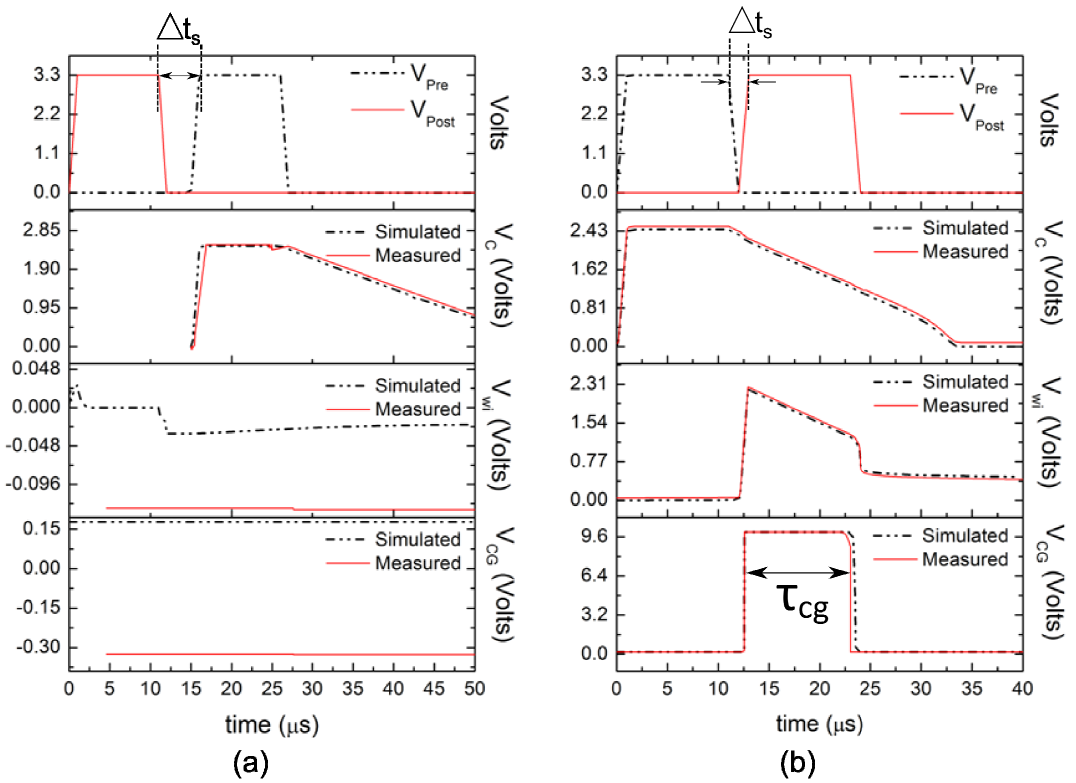


Fig. 3 – (a) Post-Pre Spiking Event - Δt_s = -5µs (b) Pre-Post Spiking Event - Δt_s = 1µs

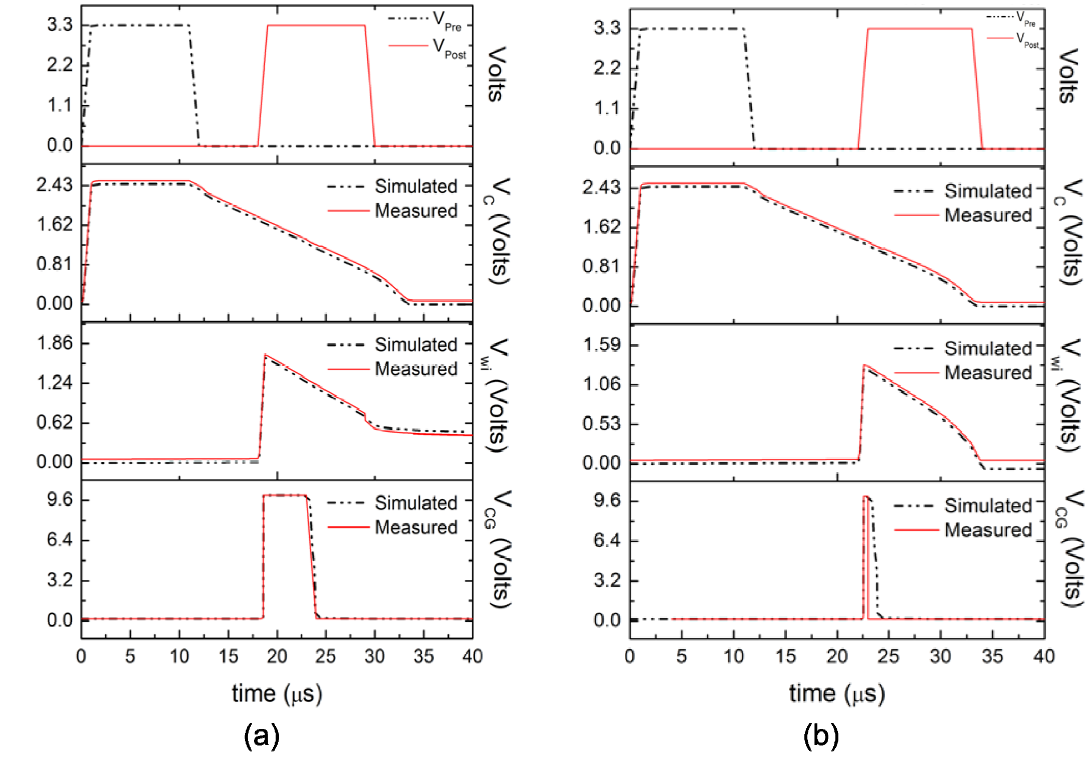


Fig. 4 (a) Pre-Post Spiking Event - Δt_s = 7µs (b) Pre-Post Spiking Event - Δt_s = 11µs

In Fig. 4(a), Δt_s is increased to 7µs, again V_{CG} is pulled high to 10V. However τ_{cg} is reduced compared to Δt_s =1µs, τ_{cg} is now 4.92µs (simulated) and 4.60µs (measured). The reduction in τ_{cg} occurs because V_{post} coincides with the linearly decreasing V_C. Voltage V_{wi} now tracks the decreasing V_C, until, eventually V_{wi} is pulled below the trigger voltage of the first CMOS inverter, while V_{post} is still high, Fig. 4(a). Finally in Fig. 4(b) Δt_s = 11µs further reduces τ_{cg} to 0.91µs and 0.65µs for simulation and measured respectively. The magnitude of V_{CG} is slightly reduced to 9.6V. This corresponds to the minimum weight update $\Delta w = \Delta w_{min}$.

Table 1 presents the results of increasing Δt_s on τ_{cg} for both simulation and experimental results. Table 1 indicates that once $\Delta t_s \geq$ 12µs then no update in the synaptic weight takes place as $V_{CG} \approx 0$ due to V_{wi} being less the threshold voltage of the first CMOS inverter when V_{post} is high. The results presented in Table 1 represented the upper left hand quadrant of the STDP curve presented later in Fig. 6.

Δt_s (µs)	τ_{cg} (µs) (Simulation)	τ_{cg} (µs) (Experimental)	V_{CG} (V)
1	10.91	10.75	10
2	9.91	9.60	10
3	8.91	8.62	10
4	7.90	7.62	10
5	6.90	6.61	10
6	5.90	5.59	10
7	4.92	4.60	10
8	3.91	3.60	10
9	2.90	2.61	10
10	1.89	1.60	10
11	0.91	0.65	9.6
12	0	0	0

Table 1 Effect of positive Δt_s on τ_{cg} and V_{CG}

III.B WD Results

As the WD circuit block is identical to the WP circuit with the exception of the application of V_{pre} and V_{post} its operation is also identical. Fig. 5(a) presents simulation and measured results of a pre-post spiking event, where the post-synaptic spike occurs 5µs after the end of the pre-synaptic spike, Δt_s = 5µs. In this case no weight update occurs. Table 2 present the simulation results for a series of post-pre spiking events upon the WD circuit. $|\Delta t_s|$ is once again increased from 1µs to 15µs. Referring to Fig. 5(b), Δt_s = -7µs; as V_{post} is pulled high C is charged to voltage V_M = 2.43V. As V_{pre} goes low, C discharges (initially) linearly via M_{leak}. When V_{pre} goes high, nodes V_C and V_{wi} are connected such that $V_{wi} \approx$ 1.70V. A weight decrease is triggered as V_{CG} is pulled down to -10V. V_{pre} goes low, both V_{wi} and V_{CG} are pulled back to 0V, ending the synaptic weight update. This is consistent with the theoretical operation outlined previously.

For Δt_s=-1µs, the maximum value of the weight decrease occurs, $\Delta w = \Delta w_{max}$. V_{CG} is at its maximum pulse width; τ_{cg} = -11.31µs and magnitude, V_{CG} = -10V. Table 2 shows that by further increasing Δt_s, to Δt_s = -5µs, Δt_s = -7µs, Δt_s = -8µs. causes τ_{cg} to be reduced to 8.14µs, 6.16µs and 5.15µs respectively. For Δt_s = -13µs $\tau_{cg} \approx$ 0.53µs, and the magnitude of V_{CG} is slightly reduced to -9.6V. This corresponds to the minimum weight update $\Delta w = \Delta w_{min}$. Table 2 indicates that once $\Delta t_s \geq$ 14µs then no update in the synaptic weight takes place as $V_{CG} \approx 0$ due to V_{wd} being less the threshold voltage of the CMOS inverter when V_{pre} is high. The results presented in Table 1 represented the lower right hand quadrant of the STDP curve presented later in Fig. 6.

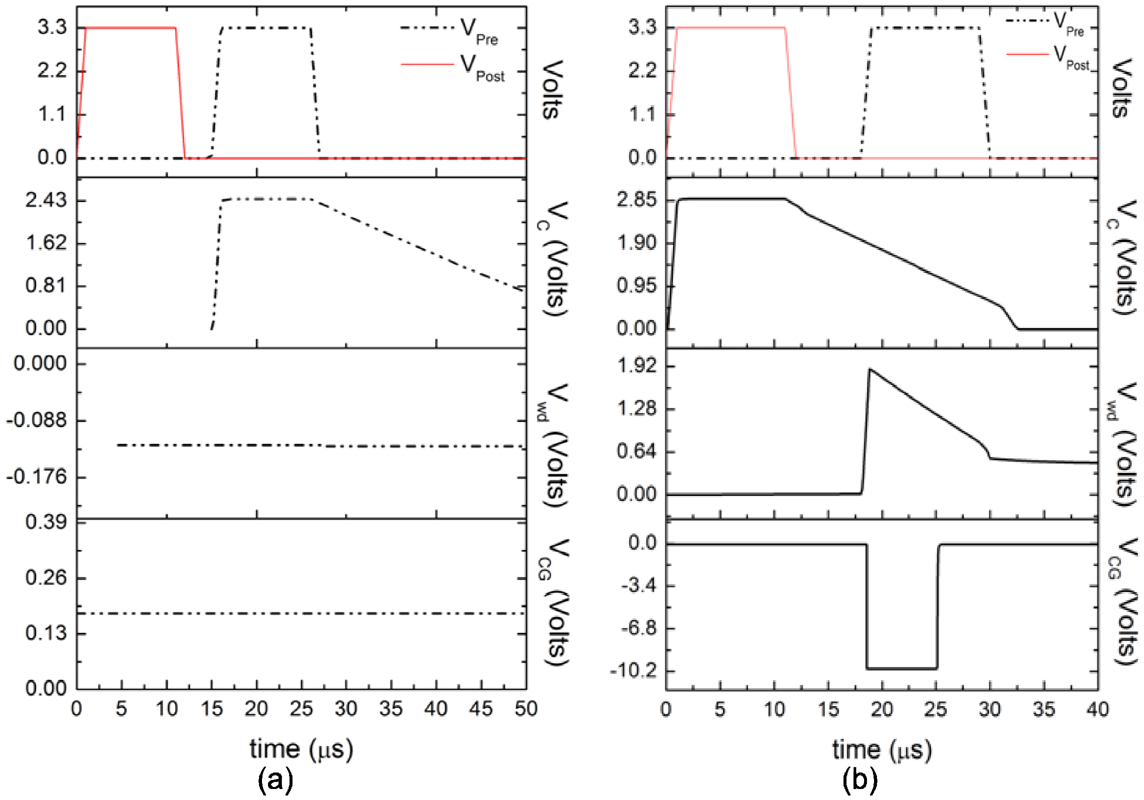


Fig. 5 – (a) Pre-Post Spiking Event - Δt_s = 5µs (b) Post-Pre Spiking Event Δt_s = -7µs

Δt_s (µs)	τ_{cg} (µs) (Simulation)	V_{CG} (V)
-1	11.31	-10
-2	10.92	-10
-3	10.18	-10
-4	9.19	-10
-5	8.14	-10
-6	7.15	-10
-7	6.16	-10
-8	5.15	-10
-9	4.14	-10
-10	3.12	-10
-11	2.06	-10
-12	0.96	-9.6
-13	0.53	-9.6
-14	0	0

Table 2 Effect of negative Δt_s on τ_{cg} and V_{CG}

Fig. 6 is a plot of τ_{cg} against Δt_s which represents the full STDP curve, shown as the insert. Note that as Δt_s is increased from 1µs to 15µs, τ_{cg} decreases from 11.31µs to ≈1µs (simulation), from 10.75µs to ≈0.65µs (measured). Similarly as Δt_s is decreased from -1µs to -15µs τ_{cg} decreases from 11.31µs to ≈0.5µs (simulation). This behaviour is characteristic of the STDP function since $\tau_{cg} \propto \Delta w$, where Q_{inj} α Δw. Note -τ_{cg} indicates a reduction in the synaptic weight.

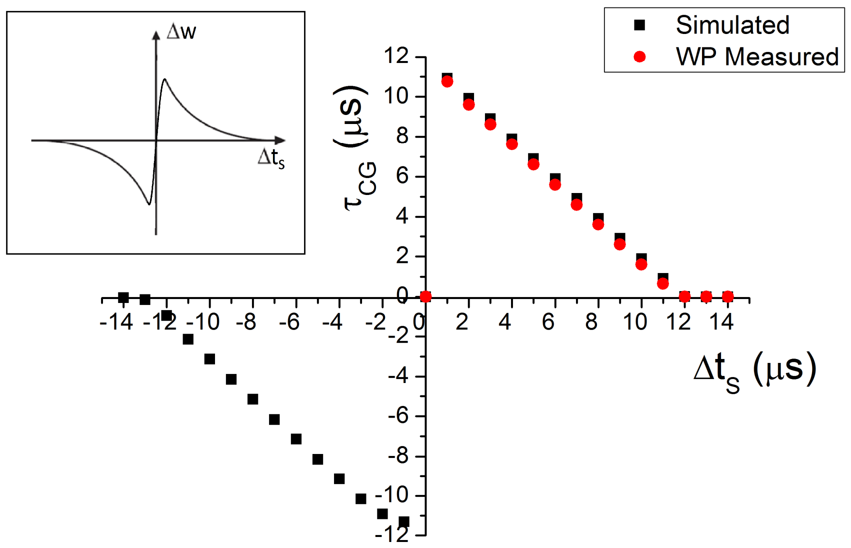


Fig. 6 STDP curve from simulation and experimental results. Insert Asymmetric STDP Curve

IV. PHYSICAL MODELLING OF WEIGHT STORAGE

The STDP circuit is to be used with FG devices, therefore we next consider the sensitivity of the weight charge injection to the FG, in relation to the STDP curve presented in Fig. 6 and charging time. The charge injected onto the FG Q_{inj} represents the change in the associated weight; Q_{inj} α Δw. The charge is injected by the Fowler-Nordheim mechanism [48].

$$J_{FN} = AE_{ox}{}^{2} exp\left(\frac{-B}{E_{ox}}\right) \qquad (5)$$

where $A = 1.54x10^{-6}\frac{m_o}{m_{ox}}\frac{1}{\phi_B}\, A/V^2$, $B = 6.83x10^7\sqrt{\frac{m_{ox}}{m_o}}\phi_B{}^{3/2}\, V/cm$, m_o is the mass of an electron at rest, m_{ox} is the effective mass of an electron in the insulator and ϕ_B is the barrier height for injection from semiconductor to oxide. It should be noted that the constants A, B are strictly for tunneling from a metal contact but are similar to the case of injection from a semiconductor [49] and serve our purpose for illustrating the model and method.

Fig. 7 presents the cross-section of a FG device constructed using a poly-silicon and MOS capacitor. The charge injected onto the FG, Q_{inj}, can be found from consideration of the current in the thin tunneling oxide, t_{ox} over a time step, Δt. We now derive a model to allow the determination of Q_{inj} (Δw) and the associated potential of charge stored on the FG, $V_{\Delta w}$.

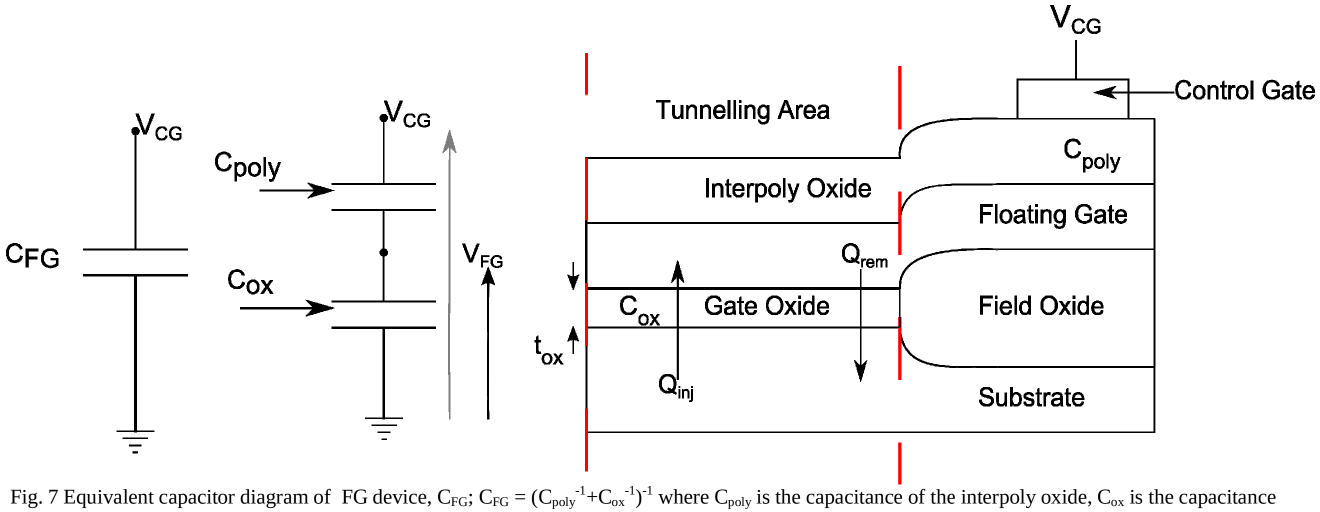


Fig. 7 Equivalent capacitor diagram of FG device, C_{FG}; $C_{FG} = (C_{poly}{}^{-1}+C_{ox}{}^{-1})^{-1}$ where C_{poly} is the capacitance of the interpoly oxide, C_{ox} is the capacitance of the tunneling oxide. V_{CG} and V_{FG} are the voltages applied to the control gate and coupled onto the FG respectively. Cross section of FG device, constructed using polysilicon and MOS capacitors. Q_{inj} represents the charge stored on the FG and Q_{rem} represents the charge removed from the FG, both due to FN tunneling.

The capacitively coupled voltage, V_{FG} which falls across t_{ox} is shown in Fig. 7, and given by $V_{FG} = \alpha V_{CG}$, where α is the capacitive coupling coefficient, defined as $\alpha = \frac{C_{poly}}{C_{ox}+C_{poly}}$. The electric field in the oxide, E_{ox} is given as $E_{ox} = \frac{V_{FG}-\phi_s}{t_{ox}}$, where it is assumed that there is no parasitic charge in the oxide or initially stored on the FG. V_{FG} is the potential of the FG and ϕ_s is the surface potential at the oxide-semiconductor interface. The field at successive time steps, Δt, can be found from equation (6) (see appendix for derivation).

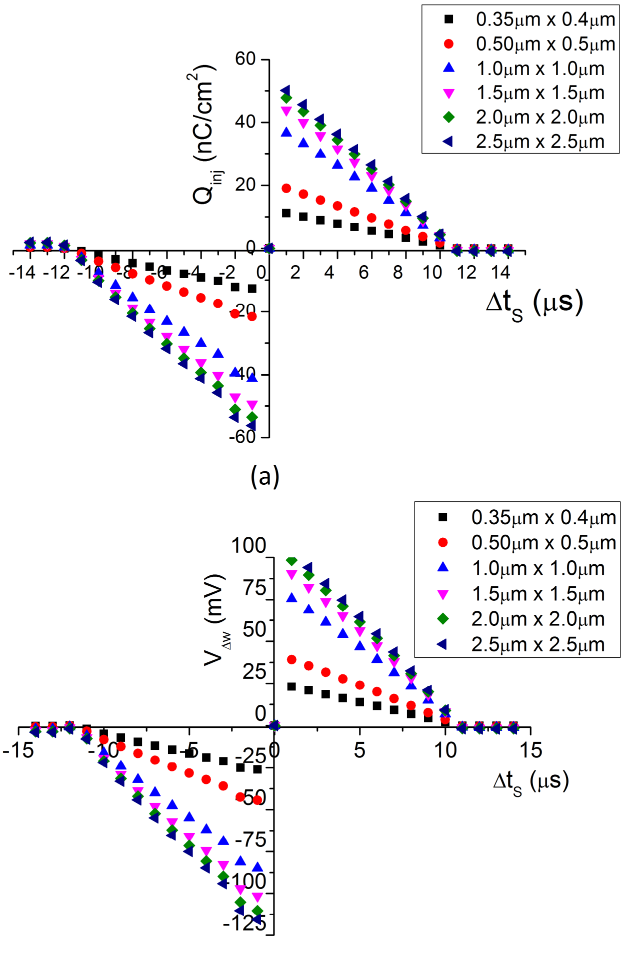
$$E_{ox(i+1)} = B\left[ln\left(\Delta t\frac{AB}{t_{ox}C_0} + exp\left(\frac{B}{E_{ox(i)}}\right)\right)\right]^{-1} \qquad (6)$$

The associated change in potential is calculated by finding the difference between successive steps of field:

$$V_{\Delta w} = t_{ox}(E_{ox}(i) - E_{ox}(i+1)) \qquad (7)$$

The charge per unit area injected onto the FG for the duration of the pulse width Δt is then found as $\Delta w \propto Q_{inj} = C_0 V_{\Delta w}$.

Fig. 8 presents plots of (a) Q_{inj} against Δt_s and (b) $V_{\Delta w}$ against Δt_s. Fig. 8 (a) presents the STDP curve for increasing tunneling area. The increment of charge injected decreases for increasing Δt because the stored charge serves to reduce the electric field.. Similarly as Δt_s is decreased below -1µs, the amount of charge removed is also decreased.



(a)

(b)

Fig. 8 STDP Curve – (a) Q_{inj} (Δw) (b) $V_{\Delta w}$

The results indicate that Q_{inj} (and $V_{\Delta w}$) tracks τ_{cg} due to the similar shape of the Q_{inj} ($V_{\Delta w}$) v Δt_s and τ_{cg} v Δt STDP plots. Increasing the device tunneling area causes a shift in the STDP curve. Specifically this is a shift in the magnitude of the charge injected/removed for the same Δt value.

The effect of process variation (PV) on the STDP curves is now considered. Fig. 9 shows the effect of PV upon the output characteristics of the STDP circuit, τ_{cg} against Δt_s. The plot concurs with the earlier statement that PV can either increase or decrease t_{cw}. The effect of this is to cause a shift in the ideal τ_{cg} against Δt_s curve. If PV causes $t_{cw} < t_{cwideal}$ (20µs) the curve is shifted to the left. Conversely if $t_{cw} > 20$µs the curve is shifted to the right.

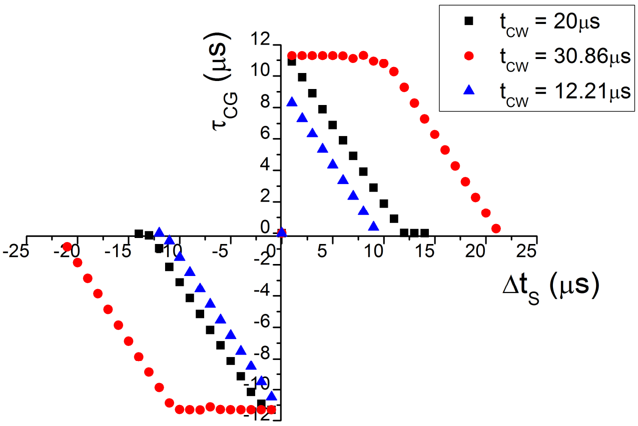


Fig. 9 τ_{cg} v Δt_s STDP curves showing effect of process variation (max, min and ideal)

The effect of PV is to vary the amount of charge (hence potential of charge) injected/removed from the FG. For $t_{cw} < 20$µs Δw ($V_{\Delta w}$) curve is shifted to the left. Conversely if $t_{cw} > 20$µs Δw ($V_{\Delta w}$) curve is shifted to the right. Specifically there is no overall change in the magnitude of Δw, Q_{inj}. Rather there is a shift in the magnitude of the charge injected/removed for the same Δt_s value. This does not affect the overall operation of the STDP circuit in that it still follows the STDP rule. However, the amount of charge injected can be compensated for by altering the learning duration.

V. Conclusion

Compact STDP circuit blocks have been proposed, which can control weight increase and decrease within a hardware neural network. Simulation and experimental results of the WP circuit are presented which indicate that for a post-pre spiking event, no update of the synaptic weight occurs. A pre-post spiking event will however cause the synaptic weight, which is represented as charge on the FG of the synapse, to be increased. The amount, by which the synaptic weight is changed, Δw, is determined by the duration that V_{wi} is greater than 1.2V and by the magnitude of V_{CG}. The maximum weight, Δw_{max} is obtained when V_{CG} has a pulse width of ≈11µs and a constant magnitude of 10V. The minimum weight, Δw_{min}, prior to V_{wi} being less than 1.2V is achieved when V_{CG} has a pulse width of 0.9µs and magnitude of 9.6V.

Furthermore, the critical timing window within which synaptic modification takes place can also be controlled with voltage, V_{leak}. The key issue of the significant influence of process variations for devices operating in subthreshold has been modeled. We show that process variations do not adversely affect the learning dynamics because the weight changes depend on the temporal difference within the STDP window. Also changes in charging/discharging duration can be compensated for within the learning algorithm. Additionally a model correlating charge alterations within the FG as a function of the charging/discharging duration was presented and this relationship was extended to show the dependency of the weight changes on the temporal difference between pre and post synaptic spikes.

Acknowledgment

The work was funded by the Engineering and Physical Sciences Research Council, UK, project reference EP/F05551X.

Appendix

Equation 6 is derived as follows.

We start with the FN Equation:

$$J_{FN} = C_0 \frac{dV_{ox}}{dt} = A{E_{ox}}^2 exp\left(\frac{-B}{E_{ox}}\right) \quad \text{(A.1)}$$

Define the time derivative of electric field as:

$$\frac{dE_{ox}}{dt} = \frac{1}{dt_{\mathrm{ox}}}\frac{dV_{ox}}{dt} \tag{A.2}$$

hence

$$J_{FN}(E_{ox}) = C_0 t_{ox}\frac{dE_{ox}}{dt} \tag{A.3}$$

Separate variables:

$$J_{FN}(E_{ox})dt = C_0 t_{ox} dE_{ox} = A{E_{ox}}^2 exp\left(\frac{-B}{E_{ox}}\right)dt \tag{A.4}$$

$$C_0 t_{ox}\frac{1}{A{E_{ox}}^2 exp\left(\frac{-B}{E_{ox}}\right)}dE_{ox} = dt \tag{A.5}$$

$$\frac{C_0 t_{ox}}{A}\int_{E_{ox}(i)}^{E_{ox(i+1)}}\left[{E_{ox}}^{-2} exp\left(\frac{B}{E_{ox}}\right)\right]dE_{ox} = \int_{t(i)}^{t(i+1)}dt \tag{A.6}$$

Where t(i+1) – t(i) = Δt, the time step. Integrating, putting in limits and re-arranging gives

$$ln\left[\Delta t\frac{AB}{C_0 t_{ox}} + exp\left(\frac{B}{E_{ox}(i)}\right)\right] = \left(\frac{B}{E_{ox(i+1)}}\right) \tag{A.7}$$

And finally,

$$E_{ox}(i+1) = B\left[ln\left(\Delta t\frac{AB}{t_{ox}C_0} + exp\left(\frac{B}{E_{ox}(i)}\right)\right)\right]^{-1} \tag{A.8}$$

REFERENCES

[1] G. Indiveri, E. Chicca and R. Douglas, “A VLSI array of low-power spiking neurons and bistable synapses with spike-timing dependent plasticity,” *IEEE. Trans. Neural Networks*, vol. 17, no. 1, pp. 211-221, 2006.

[2] C. Diorio, P. Hasler, B. A. Minch and C. A. Mead, “A single transistor silicon synapse”, *IEEE Trans. Electron Devices*, vol. 43, no. 11, pp. 1972-1980, 1996.

[3] D. H. Goldberg, G. Cauwenberghs and A. G. Andreou, “Probabilistic synaptic weighting in a reconfigurable network of VLSI integrate-and-fire neurons”, *Neural Networks*, vol. 14, pp. 781-793, 2001.

[4] L. F. Abbott and S. B. Nelson, “Synaptic plasticity: taming the beast”, *Nature Neuroscience supplement*, vol. 3, pp. 1178-1183, 2000.

[5] D.O. Hebb. *The Organisztion of Behaviour.* Wiley 1949.

[6] W.B. Levy and O. Steward, “Temporal contiguity requirements for long-term associative potentiation/depression in the hippocampus,” *Neurosience*, vol. 8, no. 4, pp. 791-797, 1983.

[7] G.Q. Bi and M.M Poo, “Synaptic modification in cultured hipocampl neurons: Dependence on spike timing, synaptic strength and postsynaptic cell type,” *J. Neuroscience*, vol. 18, pp. 10462-10472, 1993.

[8] M.Nishiyama, K. Hong, K. Mikoshiba, M.M. Poo and K. Kato, “ Calcium stores regulate the polarity and input specificity of synaptic modification,” *Nature*, vol. 408, pp. 584-588, 2000.

[9] M. Tsukada, T. Aihara, Y. Kobayashi and H. Shimazaki, “ Spatial analysis of spike-timing-dependent ltp and ltd in the ca1 area of hipocample slices using optical imaging,” *Hippocampus*, vol. 15, no. 1, pp. 104-109, 2005.

[10] H. Tanaka, T. Morie, and K. Aihara, “A CMOS spiking neural network with symmetric/asymmetric STDP function,” *IEICE Transcations on Fundamentals*, vol E92-A, no. 7, pp. 1690-1698, 2009.

[11] G.Q. Bi and M.M Poo, “Synaptic modification of corrolated activity: Hebbs postulate revisited,” *Annu. Rev. Neurosci*, vol. 24, pp. 139-166, 2001

[12] N. Caporale and Y. Dan, “Spike timing-dependent plasticity: A Hebbian learning rule,” *Annu. Rev. Neurosci*, vol. 31, pp. 25-46, 2008

[13] I. B. Levitand and L. K. Kaczmarek, *The Neuron – Cell and Molecular Biology*, 3rd Edition, Oxford University Press, 2002

[14] D. Purves, G. J. Augustine, D. Fitzpatrick,. L. C. Katz, A. LaMantina, J. O. McNamara and S. M. Willians, *Neuroscience*, 2nd Edition, Sinauer Associates Inc, 2001.

[15] N. Rebola, B. N. Srikumar and C. Mulle, “Activity-dependent synaptic plasticity of NDMA receptors”, *J. Physiol*, vol. 588, no. 1, pp. 93-99, 2010.

[16] S. Song, K. D. Miller and L. F. Abbott, “Competitive Hebbian learning through spike-timing-dependent synaptic plasticity”, *Nature Neuroscience*, vol. 3 no. 9, pp. 919-926, 2000.

[17] P. J. Dew and L. F. Abbott, “Extending the effects of spike-timing-dependent plasticity to behavioral timescales”, *PNAS*, vol. 103, no. 23, pp. 8876-8881, 2006.

[18] R. C. Froemke, D. Debanne and G. Q. Bi, “Temporal modulation of spike-timing-dependent plasticity”, *Frontiers in Synaptic Neuroscience*, vol. 2, no. 1, pp. 1-16, 2010.

[19] K. A. Buchanan, and J. R. Mellor, “The activity requirements for spike-timing-dependent plasticity in the hippocampus”, *Frontiers in Synaptic Neuroscience*, vol. 2, no. 11, pp. 1-5, 2010.

[20] Z. F. Mainen and T. J. Sejnowski, “Reliability of spike timing in neocortical neurons”, *Science*, vol. 268, pp. 1503-1506, 1995.

[21] S J. Schemmel, K. Meier and E. Mueller, “ A new VLSI model of neural microcircuits including spike timing dependent plasticity,” *IEEE International Joint Conference on Neural Networks 2004*, vol. 3, pp. 1711-1716, 2004.

[22] J. Schemmel, K. Meier and E. Mueller, “Implementing synaptic plasticity in a VLSI spiking neural network model,” *IEEE International Joint Conference on Neural Networks 2006*, pp. 1-6, 2006.

[23] K. Cameron, V. Boonsobhak, A. Murray and D. Renshaw, “Spike timing dependent plasticity (STDP) can ameliorate process variations in neuromorphic VLSI,” *IEEE Transactions on Neural Networks*, vol. 16, no. 6, pp. 1626-1637, 2005

[24] A. Bofill-i-Petit and A. F. Murray, “Synchrony detection and amplification by silicon neurons with STDP synapse,” *IEEE Transactions on Neural Networks*, vol. 15, no. 5, pp. 1296-1304, 2004

[25] Y. Hayashi, K. Saeki, and Y. Sekine, “A synaptic circuit of a pulse-type hardware neuron model with STDP,” *International Congress Series*, vol. 1301, pp. 132-135, 2007.

[26] K. Saeki, R. Shimizu and Y. Sekine, “Pulse-type hardware neural network with two time window STDP,” *ICONIP 2008, Lecture Notes In Computer Science*, vol. 5507/2009, pp. 877-884, 2009.

[27] M. M. Khan, D. R. Lester, L. A. Plana, A. Rast, X. Jin, E. Painkras and S. B. Furber, “SpiNNaker: Mapping nerual networks onto a massively-parallel chip multiprocessor,” *International Joint Conference on Neural Networks 2008,* pp.2850-2857, 2008.

[28] X. Jin, M. Lujan, L. A. Plana, S. Davies, S. Temple and S. B. Furber, “Modeling spiking neural networks on SpiNNaker,” *Computing In Science and Engineering*, vol. 12, no. 5, pp. 91-97, 2010.

[29] X. Jin, A. Rast, G. Galluppi, S. Davies, and S. B. Furber, “Implementing spike-timing-dependent plasticity on SpiNNaker neuromorphic hardware,” *World Congress on Computational Intelligence 2010*, pp. 2302-2309, 2010 Markram, H, “The blue brain project,” *Nat Rev Neurosci. vol. 7, pp. 153-160,* 2006.

[30] Druckmann, S. et al., “A Novel Multiple Objective Optimization Framework for Constraining Conductance-Based Neuron Models by Experimental Data,” *Frontiers in Neuroscience, vol. 1, no. 1, 2007*

[31] Kozloski, J. et al., “Identifying, tabulating, and analyzing contacts between branched neuron morphologies,” *IBM Journal of Research and Development, Vol 52, Number 1/2, 2008*

[32] David C. Potts, “Statistical Analog Circuit Simulation: Motivation and Implmentation”, *Advances in Analog **Circuits***, InTech, 2011.

[33] Yuhua Cheng, “The influence and modeling of process variation and device mismatch for a*nalog/RF circuit design”, Proceedings of the 4th IEEE International Caracas Conference on Devices, Circuits and Systems 2002*M

[34] .J.M. Pelgrom, A.C.J. Dunima*iker and A.P.G. Welb*^{er}*s,* “Matching Properties of MOS Transistors*”, IEEE Journal of Solid State Circuits,* ^{v}*ol. 24, no. 5, pp. 1433-1440, 1989.*

[35] M.J.M. Pelgrom, H.P. Tuinhout and M. Vertregt, “Transistor matching in analog CMOS applications”, *IEDM*, pp. 915-918, 1998.

[36] M.T. Terrovitis and C.J. Spanos, “Process Variability And Device Mismatch”, *First International Workshop on Statistical Metrology*, 1996

[37] P.G. Drennan and C.C McAndrew, “Understanding MOSFET mismatch for analog design”, *IEEE Journal of Solid State Circuits,* vol. 38, no. 3, pp. 450-456, 2003.

[38] P.R. Kinget, “Device mismatch: An analog design perspective”, *ISCAS 2007*, pp. 1245-1248, 2007.

[39] R. Jaramillo-Ramirez, J. Jaffari and M. Anis, “Variability aware design of subthreshold devices”, *ISCAS 2008*, pp. 1196-1199, 2008.

[40] H. Kosina, M. Nedjalkov and S. Selberherr, “Theory of the Monte Carlo method for semiconductor device simulation”, *IEEE Transactions on Electron Devices*, vol. 47, no. 10, pp. 1898-1908, 2000.

[41] H. . Hung and V. Adzic, “Monte Carlo simulation of device variation and mismatch in analog integrated circuits”, *NCUR 2006*, 2006.

[42] J. B. Shyu, G.C. Temes and F. Krummenacher, “Random error effects in matched MOS capacitors and current sources”, *IEEE Journal of Solid State Circuits*, vol. sc-19, no. 6, pp. 948-955, 1984.

[43] J. B. Shyu, G.C. Temes and K. Yao, “Random error in MOS capacitors”, *IEEE Journal of Solid State Circuits*, vol. sc-17, no. 6, pp. 1070-1076, 1982

[44] B. Zhai, S. Hanson, D. Blaauw and D. Sylvester, “Analysis and mitigation of variability in subthreshold design”, *ISLPED 2005*, pp. 20-25, 2005

[45] S. N. Mozaffari and A. Afzali-Kusha, “Statistical model for subthreshold current considering process variation”, *ASQED 2010*, pp. 356-360, 2010

[46] R. Kempter, W. Gerstner and J.L. van Hemmen, “Hebbian learning and spiking neurons”, *Phys. Rev. E*, vol. 59, pp. 4498-4514, 1999.

[47] W. Gerstner, R. Kempter, J. L. van Hemmen, and H. Wagner, “A neuronal learning rule for sub-millisecond temporal coding”, *Nature*, vol. 386, pp. 76-78, 1996.

[48] R.H. Fowler and L. Nordheim, “Electron emission in intense electric fields”, *Proceedings of the Royal Society of London **A***, vol. 119, pp. 173-181, 1928

[49] Z. A. Wienberg, “On tunneling in metal-oxide-silicon structures”, Journal of Applied Physics, vol. 53, no. 7, pp. 5052-5056, 1962

[50] P. D. Roberts and C. C. Bell, “Spike timing dependent synaptic plasticity in biological systems”, *Biological Cybernetics*, vol. 87, no. 5-6, pp. 392-403, 2002.

[51] B. Lu, W.M. Yamada, and T. W. Berger, “Asymmetric Synaptic Plasticity Based on Arbitrary Pre- and Postsynaptic Timing Spikes Using Finite State Model”, *Proceedings of International Joint Conference on Neural Networks*, Orlando, Florida, USA, August 12-17, 2007

[52] T. Dowrick, S. Hall and L. McDaid, “A silicon based dynamic synapse with depressing response”, *IEEE Transactions on Neural Networks and Learning Systems,* Vol.23, no. 10, pp. 1513-1525, 2012.