Neurocomputing 124 (2014) 210–217

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A compact spike-timing-dependent-plasticity circuit for floating gate weight implementation

A.W. Smith, L.J. McDaid, S. Hall *

University of Liverpool, Department of Electrical Engineering & Electronics, Brownlow Hill, Liverpool L69 3GJ, United Kingdom

ARTICLE INFO

Article history:
Received 16 April 2012
Received in revised form
16 July 2013
Accepted 23 July 2013
Communicated by R.W. Newcomb
Available online 30 August 2013

Keywords:
Spike timing dependent plasticity
Neural networks
Floating gate
MOSFET

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 analyzed. 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 from 10 μs to 100 ms 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.

1. Introduction

Significant research over the last 2 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 [7,13–19] typically occurs over the range 10–100 ms 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 four MOS transistors, and a polysilicon capacitor.

* Corresponding author. Tel.: +44 151 794 4529; fax: +44 151 794 4540.
E-mail address: s.hall@liverpool.ac.uk (S. Hall).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.07.007

The paper is organized as follows: in Section 2 an overview of theoretical operation of the compact STDP circuit is presented. Section 3 presents experimental and simulation results undertaken in AMS 0.35 μm 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 4 and conclusions drawn in Section 5.

2. 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.

2.1. 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 .

The WP and WD circuits each consist of three NMOSTs, $\mathrm{M_{Pre}}$, $\mathrm{M_{Post}}$ and $\mathrm{M_{leak}}$, a PMOST, $\mathrm{M_{reset}}$ and a MOS capacitor, C. Transistor $\mathrm{M_{reset}}$ 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, $\mathrm{M_{reset}}$ is off and will not significantly affect V_{wi} or V_{wd}. 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 $\mathrm{M_{leak}}$ 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 $\mathrm{M_{pre}}$. When the pre-synaptic pulse ends, C starts to discharge via $\mathrm{M_{leak}}$, and $\mathrm{V_C}$ 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 $\mathrm{M_{post}}$. 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 3.3 V and 10 V 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 10 V. 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 $\mathrm{V_{CG}}$ determines how much charge is injected and stored on the FG. As $\Delta t_\mathbf{s} \rightarrow \Delta t_{\mathbf{s}\ \min}$, $\tau_{cg} \rightarrow \max\ \tau_{cg}$. Similarly as $\Delta t_\mathbf{s} \rightarrow \Delta t_{\mathbf{s}\ \mathrm{man}}$, $\tau_{cg} \rightarrow \min\ \tau_{cg}$. 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 3.3 V and -10 V 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 -10 V. If V_{wd} is less than the threshold voltage of the inverter, then the output is 0 V. 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 $\Delta t_\mathbf{s}=0$ (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 0 V. 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 ($\Delta t_\mathbf{s}=0$).

2.2. 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 20–25 ms 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 20–25 μs 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 $\mathrm{M_{leak}}$, S_{Mleak}. The sub-threshold current, I_{leak} is constant for $V_{DS}=V_C>3$ kT/q;

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

where V_t is the threshold voltage of $\mathrm{M_{leak}}$, q is the charge of an electron, k is the Boltzmann constant and T is absolute temperature. The sub-threshold slope parameter, $m=1+\mathrm{C_d}/\mathrm{C_o}$ with $\mathrm{C_d}$ being the depletion layer capacitance, $\mathrm{C_o}$ 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

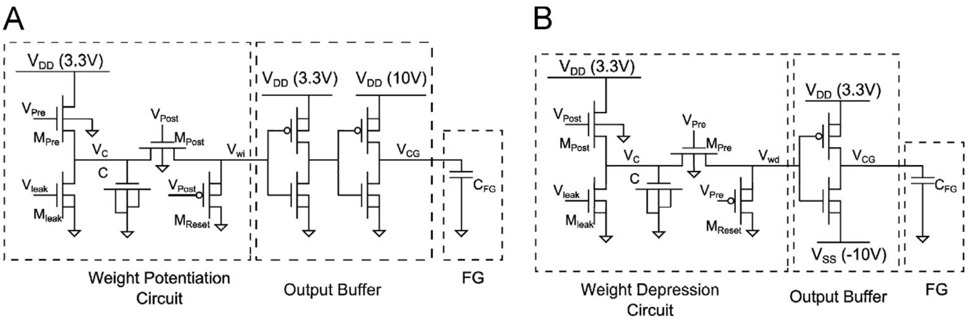

Fig. 1. (a) WP and (b) WD circuit block with FG device and driver buffer circuit. Voltages indicated are relative to ground.