Heterogeneous Neurons and Plastic Synapses in a Reconfigurable Cortical Neural Network IC

Jayawan H B Wijekoon and Piotr Dudek

School of Electrical and Electronic Engineering,
University of Manchester, UK

jayawan@ieee.org; p.dudek@manchester.ac.uk

Abstract— **This paper presents an analogue VLSI circuit intended to be used in a neural network architecture that closely resembles the small-scale laminar micro-circuits of the neocortex. The Cortical Neural Layer (CNL) chip comprises of 120 reconfigurable cortical neurons and 7,560 synapses. The neurons can be configured to produce regular spiking, fast spiking, chattering, intrinsically bursting, and other complex activity patterns. The synaptic circuits include inhibitory/ excitatory, facilitating/depressing and spike-time dependent plasticity (STDP) dynamics. The connectivity of the neural network can be configured using off-chip spike-routing and on-chip axonal arbor connections. A pre-synaptic spike can be sent to a group of crossbar synapses simultaneously, reducing latency in the pre-synaptic spike routing, enabling a high degree of connectivity of the neural network. The device is fabricated in a 0.35 µm CMOS technology and on-chip neural dynamics are experimentally verified.**

I. Introduction

The primate brain performs sophisticated functions, with remarkably low energy consumption. Of all the brain regions, the neocortex is known to be essential for psychophysical signal processing and the generation of intelligent behaviour. In parallel with the research effort to understand the principles of information processing of the neocortex, through simulating neuronal networks on conventional computers [1], the research on neuromorphic devices aims to provide compact, real- or accelerated- time, low-power brain-like circuits in VLSI hardware [2-5]. It is hoped that the design and implementation of these devices will pave the way towards a realisation of portable, low-power, intelligent computing systems.

The brain consists of highly specialised structures that are built with complex, heterogeneous neural elements. Individual neurons or synapses show diverse non-linear responses to the same inputs. At present, it is not known which of the cortical neuron cell behaviours, synaptic dynamics and plasticity mechanisms are the essential aspects of neural computation. Most of the neuromorphic circuit implementations use relatively simple models such as integrate and fire point neurons with homogenous neural responses, fixed weight synapses, and simplifying assumptions about the connectivity. This allows larger networks to be implemented on a single integrated circuit (IC), however, biological plausibility is sacrificed.

In this paper, we present the Cortical Neural Layer (CNL) integrated circuit that consists of 120 spiking and bursting neurons and 7,560 synapses with an assortment of short- and long- term plasticity mechanisms, including facilitation, depression and spike-timing dependent plasticity (STDP). We aimed to maintain the circuit simplicity that would allow a high level of integration, but at the same time provide an increased level of biological plausibility. The established computational models [6-8] are used as a guide to arrive at circuit models that demonstrate qualitatively similar behaviors. The circuits are designed with a degree of generality such that the externally adjustable parameters can be tuned to obtain distinct classes of neural responses, providing a flexible and versatile tool for simulating cortical circuits. The circuits operate at approximately three orders of magnitude faster than biology. The CNL is fabricated in a standard 0.35 µm CMOS technology and fits in a 24 mm^2 (6.78 mm by 3.58 mm) footprint.

II. Chip Architecture

The neural elements in the CNL chip occupy two separate blocks, Block-A and Block-B, as shown in Fig. 1. Block-A consists of 100 neurons and 63×100 synapses. Each neuron in this block receives inputs from 21 excitatory STDP synapses, 22 non-STDP excitatory depressing (ED) synapses including an internally connected non-crossbar synapse), and 20 inhibitory depressing (ID) non-STDP synapses (3 somatic and 17 distal inhibitory synapses; these two types are distinguished by the range of available synaptic strengths). Block-B consists of 20 neurons and 63×20 synapses. Each Block-B neuron receives inputs from 63 non-STDP synapses. These comprise of an equal number of excitatory facilitating (EF), inhibitory facilitating (IF), and excitatory depressing (ED) synapses.

About 80% of the neurons in a cortical network of neocortex are excitatory, and others are inhibitory. Anatomically, these two types of neuron are equipped with different types of input and output synapse combinations. Block-A and Block–B in the CNL chip are designed to represent excitatory and inhibitory populations of neurons respectively. The variety of synapse types and the percentage of each synapse type in a Block are chosen according to the anatomy of the neocortex. Hence, if the output of a Block-A neuron is connected to excitatory synapses, such a neuron represents an excitatory neuron of a cortical network. Similarly, if the output of a Block-B neuron is connected to inhibitory synapses, and with an appropriate input synapse combination, such a neuron represents an inhibitory neuron of a cortical network. The 120 neurons are grouped into 13 groups (made up of different numbers of neurons) where each group can be independently configured to a particular neuron type- this allows variety of neuron type groups in each Block.

The pre-synaptic input spikes are provided externally, by sending an address of the synapse as an address event. The event address comprises of a row and column address, where every address bit can have a '*don't care*' value (in practice, this is done by using two bits to code '0', '1' and '*don't care*'). This allows pre-synaptic spike events to be

978-1-4673-0219-7/12/$31.00 ©2012 IEEE

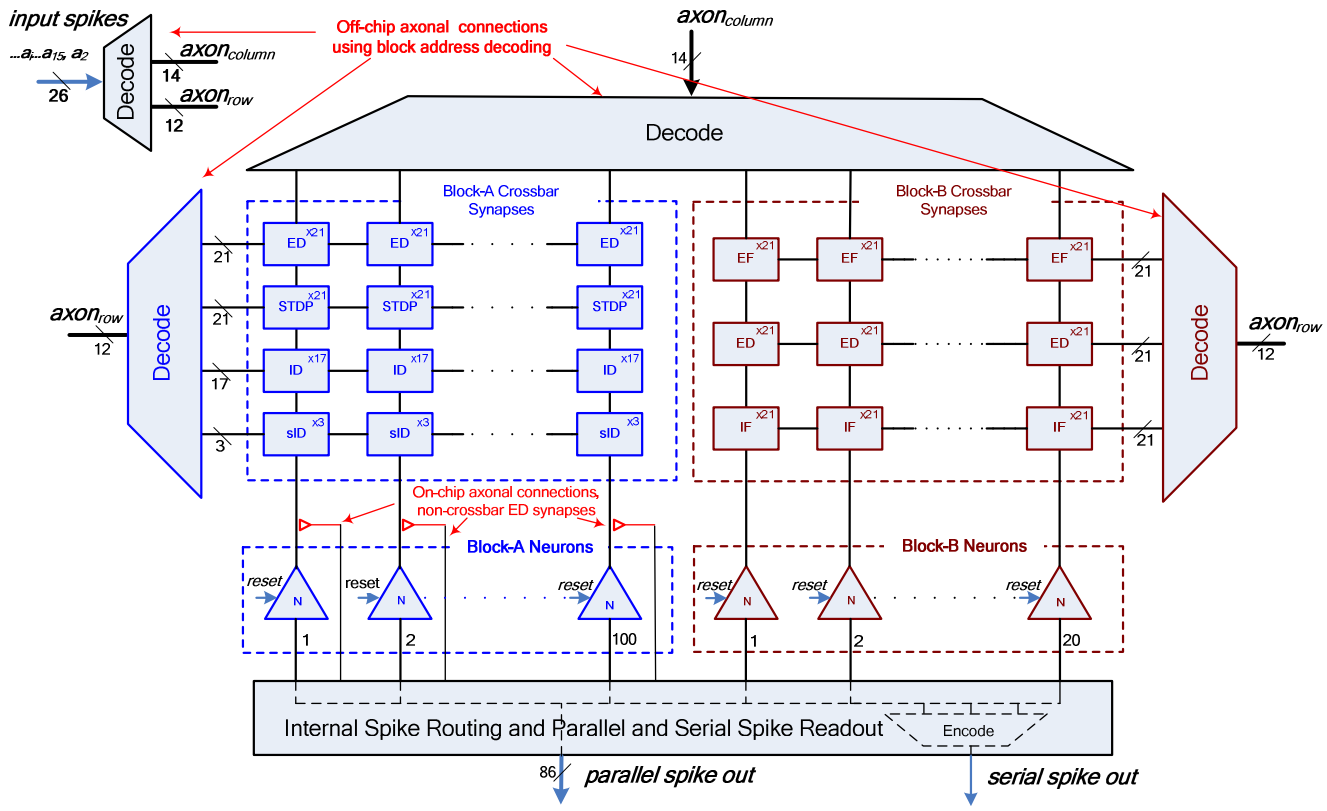

Fig. 1. Main internal circuit blocks of the CNL core; Synaptic circuits are of the following types: ED-Excitatory Depressing, STDP- Spike-Time Dependent Plasticity, ID- Inhibitory Depressing, sID-somatic Inhibitory Depressing, EF- Excitatory Facilitating, IF- Inhibitory Facilitating, N-Neuron.

sent to all or to a selected group of synapses simultaneously. Eighty six of the neuron outputs are available in parallel from the chip pins, sixteen of the neurons are internally routed and the remaining neuron outputs are accessible serially. The connectivity matrix, the spike routing of the network and axonal/dendritic delays can be set up using an external programmable device such as a Field Programmable Gate Array (FPGA).

There are also some on-chip routed axonal connections: the outputs of eight Block-A neurons project to 32 other neurons in Block-A via additional ED synapses (each output projects to four consecutive neurons); the outputs of eight Block-B neurons project to 64 other neurons in Block-A, via additional ED synapses (each output projects to eight Block-A neurons). These hard-wired axonal arbor connections can be switched -off or -on when required.

The resting weights of the synapses are set externally so that the synapses of one type (ED, ID, IF or EF) can have a linear distribution of weights along the column. Two parameters are used to set the highest and the lowest resting weight in the group, per synapse type. These weight distributions are common to all columns/neurons. Each synapse type has one set of externally controllable tuning parameters.

III. Core Neural Circuits

A. Cortical Neuron Circuit

In the CNL chip we have used our compact generic cortical neuron circuit [9]. The model is inspired by the computational model proposed in [8], and the circuit can be configured to represent a variety of cortical neuron cells such as regular spiking, chattering, intrinsic bursting, fast spiking, etc.

B. STDP Synapse Circuit

The STDP mechanism is thought to be involved in learning, memory and cortical plasticity [7]. If the post- synaptic neuron fires, the input synapses could reduce (long-term depression, LTD) or increase (long-term potentiation, LTP) the strength of the synaptic transmission. The amount of depression or facilitation depends on the time difference between the pre- and post- synaptic spikes. Experimentally observed STDP effects are diverse and depend on the synapse and neuron type [10]. The STDP circuit on the CNL chip produces a family of STDP curves similar to the one shown in Fig.2, which was obtained experimentally.

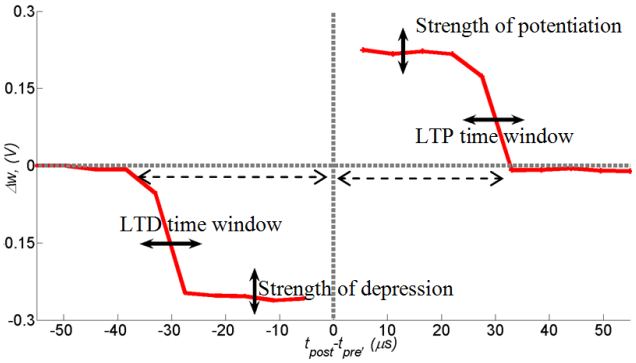

Fig. 2. The shape of the STDP curve implemented on the chip; the time window of LTP and LTD and the strength of potentiation and depression can be tuned using the parameters of the synapse circuit.

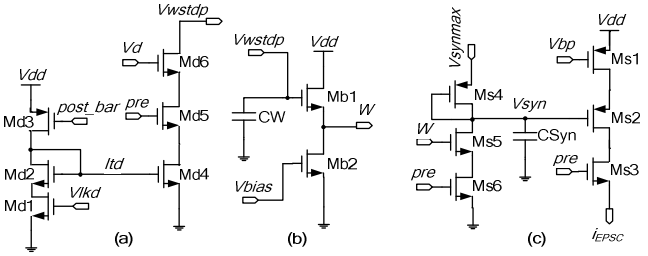

Fig. 3. STDP synapse circuit: (a) LTD, (b) WBUF, (c) STDP-ISYN. The synapse also includes an LTP circuit, complementary to the LTD circuit.

The schematic diagram of the STDP synapse circuit is shown in Fig.3. The circuit comprises of 4 sub-circuits: Long-Term Depression (LTD) circuit, Long-Term Potentiation (LTP) circuit, Weight Buffer (WBUF) circuit, and post-synaptic current generator (STDP-ISYN) circuit. The LTD and LTP circuit topologies were initially proposed by Indiveri [11] and operate on a biological time scale. By appropriately sizing the transistors and capacitors and by shifting the operating point of some transistors, the circuits are designed to operate on an accelerated time scale. The weight of the synapse, *Vwstdp* is stored on the capacitor C_W. Firings of the pre- and post- synaptic neurons induce changes in the synaptic weight, using the LTD and/or LTP circuits to implement the STDP rule. The signals labeled *pre, post* and *post_bar* in the circuit schematics are the pre-synaptic firing signal, post-synaptic firing signal, and the inverted post-synaptic firing signal respectively. Once the post-synaptic neuron fires, the gate capacitance of Md2 is charged to the supply voltage, *Vdd,* by switching on the transistor Md3. The gate capacitance is then continuously discharged with a "leakage" current, through transistors Md1 and Md2, the amount discharged is approximately proportional to the time after the last post-synaptic spike. The time duration for the capacitor to discharge to a voltage that is low enough to force the gate voltage of the transistor Md4 to reach its cut-off region of operation is equivalent to the '*LTD time window'*. This is controlled by the voltage *Vlkd*. If a pre-synaptic spike follows the post-synaptic firing within the LTD time window, the LTD circuit reduces the charge in the capacitor C_W by switching on the current path through transistors Md4, Md5 and Md6. The voltage *Vd* limits the maximum current through these transistors. The LTP circuit is complementary to the LTD circuit, the *ltp* node is discharged by the pre-synaptic spike and slowly charged up (at a rate controlled by voltage *Vlkp*), so that an amount of charge (dependent on the time between the pre- and post-synaptic spike) is transferred onto C_W when the post-synaptic spike arrives.

The WBUF circuit buffers the synaptic weight and provides it to the STDP-ISYN circuit, to generate excitatory post-synaptic current within the desired range for all possible values of the buffered weight. Here, the post-synaptic current, i_{EPSC} is injected in the form of a current impulse (i.e. rise- and fall- time constants of the post- synaptic potential are not considered). The i_{EPSC} current can be scaled or limited using an externally controllable voltage *Vbp*. The circuit holds the synaptic weight using the capacitor C_W; the synaptic weight value leaks with approximately a 40 ms average time constant. All times are scaled by 10^{-3} with respect to biological time.

C. Short-Term Plastic Synapse Circuit

The synaptic facilitating and depressing circuit designs are based on the computational neural model of short-term dynamics, proposed in [6]. The schematic diagram of the ED synapse is shown in Fig. 4. The WSYN circuit in

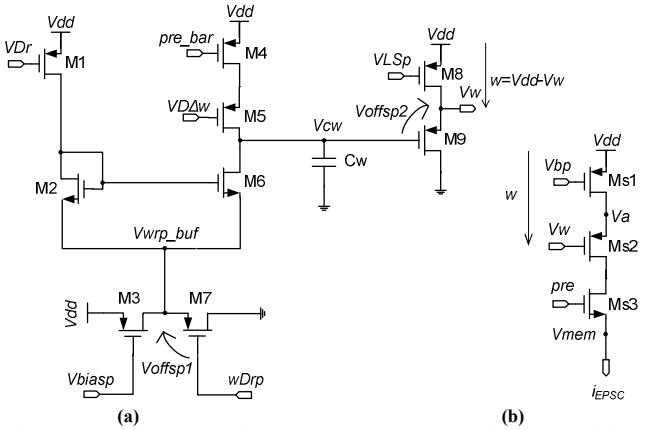

Fig. 4 ED short-term dynamic synapse: (a) Excitatory Weight Depressing (WSYN) circuit, (b) EX-ISYN circuit.

Fig.4 (a) produces short-term depressing dynamics and passes the synaptic weight to the EX-ISYN circuit which generates a post-synaptic current when a pre-synaptic spike arrives. The weight of the synapse is represented by the voltage at the node *Vw* with reference to *Vdd*. It can have a value between the resting weight of the synapse and zero, depending on the short-term neural activity of the pre-synaptic neuron. The *Vw* follows the voltage across the capacitor Cw with an off-set (*Voffsp2*) produced by the source follower buffer (transistors M8-M9).

During the pre-synaptic neuron spike, finite charge is added to the capacitor Cw through the transistors M4 and M5. Consequently, the weight is depressed. The voltage *VDΔw* controls the amount of depression. Depending on the chosen level of *VDΔw*, the amount of depression can also depend on the weight of the synapse (when the transistor M5 is in the linear region). The current mirror circuit (transistors M1, M2 and M6) continuously discharges the capacitor towards the voltage *Vwrp_buf*. The source follower circuit (transistors M3 and M7) buffers the voltage *wDrp* onto the *Vwrp_buf* node. Hence, the resting weight can be set by the externally controlled bias voltage *wDrp*. The rate of discharge towards the resting weight (i.e. the speed of recovery of the depressing synapse) is controlled by voltage *VDr*. By setting *VDΔw* to *Vdd*, the depressing synaptic dynamics can be switched off completely, and the synapse can be used as a simple excitatory synapse. The *Vbiasp* and *VLSp* are used to bias the source follower circuits.

When the pre-synaptic spike arrives at the gate of transistor Ms3 of the EX-ISYN circuit shown in Fig. 4 (b), the circuit generates an excitatory post-synaptic current (i_{EPSC}) approximately proportional to the square of the synaptic weight. The i_{EPSC} current for a given value of weight can be scaled or limited using the externally controllable voltage *Vbp*, depending on the operational region of Ms1.

The other synapse circuits (ID, EF, IF) follow a similar design, using the WSYN and EX-ISYN circuits, or their complementary forms, and level-shifting circuits as required.

IV. Experimental Results

The CNL chip and Xilinx Virtex 5 FPGA (used to interface & route the input/output spikes) are placed on a custom printed circuit board. The board is interfaced to a PC through a USB interface, which is used to configure the device and to read the spiking activities of the network. The tuning parameters that define the characteristics of the synapses and neurons are provided from the PC, through a programmable analogue voltage output PCI card, and the

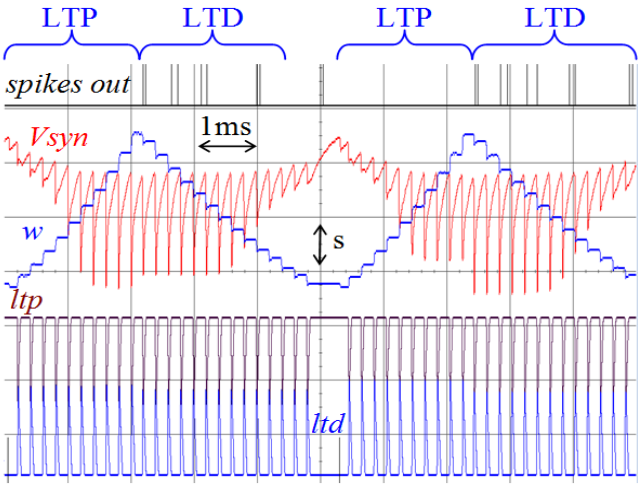

Fig. 5. Experimental waveforms of the fabricated STDP synapse circuit: waveforms include the node voltages of the circuit shown in Fig. 3, *Vsyn*, *w*, *ltp* and *ltd*, and *spikes out* (output of the neuron) in response to twenty four pre- and post- synaptic spike pairs - initially, the post-synaptic spike (indicated by a step increase in *ltd* trace) follows the pre-synaptic spike (indicated by a step decrease in *ltp* trace) and then the pre-synaptic spike follows the post-synaptic spike. Voltage scale, s, for the node voltages, *Vsyn*, *w*, *ltp*, and *ltd* is 0.05V, 0.2V, 1V, and 1V respectively. Here, to test the function of the STDP circuit, artificial post-synaptic spikes are generated for the STDP circuit from the FPGA, by disconnecting post-synaptic input from the post-synaptic neuron.

internal states of selected synapses are observed on an oscilloscope.

The behavior of STDP synapses is experimentally demonstrated in Fig. 5. The waveforms include traces of the node voltages of the circuit in Fig. 3: *Vsyn*, *w*, *ltd*, and *ltp* (from the complementary LTP circuit) and the output spikes of the post-synaptic neuron. Experimental results from the fabricated short-term plastic synapse circuits: ED, ID, EF, and IF synapses are shown in Fig. 6. The circuits behave as expected, producing the required functional behaviours implementing the short- and long- term plasticity rules. The

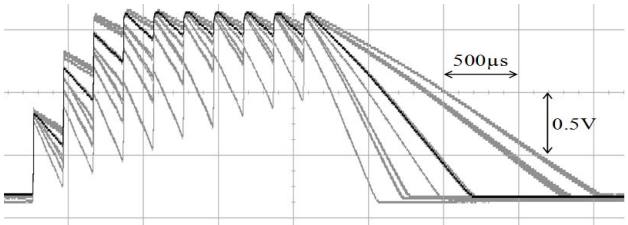

Fig. 7. The synaptic weight traces of the fabricated IF synapse circuits across eight synapses, in response to to the same stimulus (the burst of pre-synaptic spikes followed by a silent period) and with identical bias/tuning voltages.

presented waveforms are for a typical set of tuning parameters, which are widely adjustable. To demonstrate the degree of repeatability, the results in Fig.5 and Fig.6 are shown for two iterations of the same stimuli. Variation of the waveforms of the IF synaptic weight, across eight synapses that use the same tuning parameters, are shown in Fig. 7 to demonstrate the degree of mismatch variation of the fabricated short-term dynamic synapse circuits.

V. Conclusions

The Cortical Neural Layer integrated circuit consists of reconfigurable cortical neurons, and synapses with short-term dynamics and STDP plasticity mechanisms. The functionalities of these neural elements are experimentally verified. The circuits can be used to construct neural

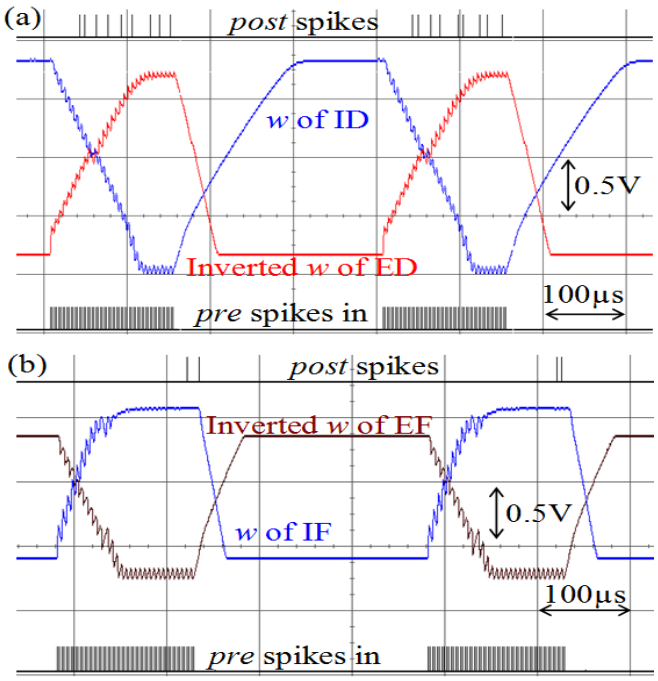

Fig. 6. Experimental results from the fabricated short-term plastic synapse circuits, demonstrating the short-term facilitation and depression of the synapse weight, *w*, for the burst of pre-synaptic spikes followed by a silent period; (a) Block-A Excitatory Depressing (ED) and Inhibitory Depressing (ID) synapse's weight response, (b) Block-B Excitatory Facilitating (EF) and Inhibitory Facilitating (IF) synapse's weight response.

networks that closely approximate the micro-circuits of the neocortex.

References

[1] Markram H., "The Blue Brain Project", Nature Review Neuroscience, 7(2), pp.153-160, 2006.

[2] Vogelstein R. J., et.al. "Dynamically Reconfigurable Silicon Array of Spiking Neurons With Conductance-Based Synapses", IEEE Transactions on Neural Networks, 18(1), pp.253-265, 2007.

[3] Giulioni M. et. al. "A VLSI network of spiking neurons with plastic fully configurable "stop-learning" synapses", IEEE International Conference on Electronics, Circuits and Systems, ICECS 2008 678-681, IEEE (Eds.), 2008.

[4] Schemmel J., Fieres J., and Meier K., "Wafer-Scale Integration of Analog Neural Networks" IEEE International Joint Conference on Neural Networks, 1-8 June, pp.431 – 438, 2008.

[5] Merolla P. et.al. "A Digital Neurosynaptic Core Using Embedded Crossbar Memory with 45pJ per Spike in 45nm", Proceedings of the IEEE Custom Integrated Circuits Conference, 2011.

[6] Abbott L. F., Varela J. A., Sen K., Nelson S. B., "Synaptic depression and cortical gain control", Science, 275(5297), pp.221-224,1997.

[7] Tsodyks M., "Spike-timing-dependent synaptic plasticity - the long road towards understanding neuronal mechanisms of learning and memory", Trends in Neurosciences, 25(12), pp. 599-600, 2002.

[8] Izhikevich E. M., "Simple model of spiking neurons", IEEE Transactions on Neural Networks, 14(6), pp.1569-72, 2003.

[9] Wijekoon J. H. B. and Dudek P., "Compact Silicon Neuron Circuit with Spiking and Bursting Behaviour", Neural Networks, 21 (2-3), pp.524-534, March/April 2008.

[10] Abbott L. F., Nelson S. B. , "Synaptic plasticity: taming the beast", Nat. Neurosci., pp.1178–1183, 2000.

[11] Indiveri G., "Neuromorphic bistable VLSI synapses with spike-timing-dependent plasticity", In Advances in Neural Information Processing Systems,2003.