Modeling Synaptic Plasticity within Networks of Highly Accelerated I&F Neurons

Johannes Schemmel, Daniel Brüderle, Karlheinz Meier and Boris Ostendorf
Kirchhoff Institute for Physics, University of Heidelberg
Im Neuenheimer Feld 227, 69120 Heidelberg, Germany
Email: schemmel@kip.uni-heidelberg.de

***Abstract*— When studying the different aspects of synaptic plasticity, the timescales involved range from milliseconds to minutes, thus covering at least seven orders of magnitude. To make this temporal dynamic range accessible to the experimentalist, we have developed a highly accelerated analog VLSI model of leaky integrate and fire neurons. It incorporates fast and slow synaptic facilitation and depression mechanisms in its conductance based synapses. By using a 180 nm process 10^5 synapses fit on a 25 mm^2 die. A single chip can model the temporal evolution of the synaptic weights in networks of up to 384 neurons with an acceleration factor of 10^5 while recording the neural action potentials with a temporal resolution better than 30 μs biological time. This reduces the time needed for a 10 minute experiment to merely 6 ms, paving the way for complex parameter searches to reproduce biological findings. Due to a digital communication structure larger networks can be built from multiple chips while retaining an acceleration factor of a least 10^4.**

I. Introduction

In recent years a new interest in implementing biologically realistic neural systems based on spiking neurons could be observed [1][2]. Part of the motivation lies in the increased consensus within the neuroscience community considering single neuron models and the modeling of synaptic transmission including the different aspects of plasticity.

The VLSI model presented in this paper is part of this ongoing approach. It differs from the cited examples insofar as its tries to achieve an acceleration factor as high as possible compared to biological time. Together with the high number of provided synapses—the presented system can be scaled up to more than 10^6 synapses—it aims to complement numerical simulations. One aspect would be the statistical analysis of the temporal development of synaptic transmission due to plasticity. With a speed-up factor of 10^5 for a single chip and still more than 10^4 for networks built of multiple dies it is possible to do extensive parameter searches even for experiments spanning several minutes of biological time.

The chosen neuron model is the standard leaky integrate and fire with conductance based synapses [3]. To facilitate the communication between the neurons, the action potential (AP) is propagated as a digital pulse. Conductance-based synapses connect these digital neuron outputs to the membranes of other neurons. In the presented chip 256 synapses connect to one neuron, a number limited by the size of the chip.

Two plasticity mechanisms are implemented in the synapse circuits which are both based on the temporal change of synaptic transmission. Although they modulate the same parameter the timescales involved differ by more than four orders of magnitude. Short term plasticity is based on the history of pre-synaptic APs. It emulates the limitation of resources involved in the synaptic transmission, like for example neurotransmitters [4]. The temporal evolution of the network is caused by long term synaptic plasticity based on *spike time dependent plasticity* (STDP) [5][6]. In this model each synapse measures the correlation between pre- and post-synaptic APs which is then used to calculate long term changes in the synaptic weights. While synaptic plasticity is implemented on the circuit level in the presented system, slower adaptation processes—like neuro-modulators—as well as developmental changes in neuronal connectivity can be easily incorporated into the digital control of the analog continuous-time model.

II. Utilized Models

A. Neuron and Synapse Model

The membrane potential $V(t)$ is governed by the following differential equation:

$$C_m \frac{dV}{dt} = g_m(V - E_l) \quad + \sum\nolimits_k p_k(t) g_k(t)(V - E_x) \\ + \sum\nolimits_l p_l(t) g_l(t)(V - E_i) \qquad (1)$$

Each term on the right hand side contributes an individual current to the total membrane current, which by itself is equal to the derivative of the membrane potential multiplied by a constant C_m representing the total membrane capacitance. The first term models the contribution of the different ion channels that determine the potential E_l the membrane will eventually reach if no other currents are present. The synapses use different reversal potentials, E_i and E_x, to model inhibitory and excitatory ion channels. The index k in the first sum runs over all excitatory synapses while the index l in the second covers the inhibitory ones. The individual activations of the synapses are controlled by the *synaptic open probability* $p_{k,l}(t)$ [3]. Plasticity is included in the model by varying g_k and g_l with time. Since the involved timescales vary greatly between short-term and long-term plasticity, both mechanisms act at different stages of the synaptic signal transmission chain. The synaptic conductance $g_{k,l}$ is modeled as a product of the synaptic weight $\omega_{k,l}(t)$ and a maximum conductance $g_{\max k,l}(t)$.

$$g_{k,l}(t) = \omega_{k,l}(t) \cdot g_{\max k,l}(t) \qquad (2)$$

The weights are used for the initial setup of the connection strengths. They are modified by the implemented long-term plasticity algorithm (STDP) and thus vary slowly with time t. Short-term plasticity acts by temporarily increasing or decreasing the maximum conductance $g_{\max k,l}(t)$.

B. Short Term Synaptic Plasticity

The implemented model follows the ideas developed in [4][7]. In the case of short term depression the *absolute synaptic efficacy* A_{SE} is thought to be distributed between an inactive (I) and a recovered partition (R). With each AP a conductance pulse with $g_{\max}$ proportional to the percentage of the total efficacy momentarily in the recovered partition is generated. After the AP was communicated to the post-synaptic neuron a fixed fraction of the recovered efficacy, the *usable synaptic efficacy* U_{SE}, is transferred to the inactive partition. While this transfer repeats with each AP the inactive partition loses efficacy to the recovery

This work is supported in part by the European Union under the grants no. IST-2001-34712 (Sensemaker) and no. IST-2005-15879 (FACETS).

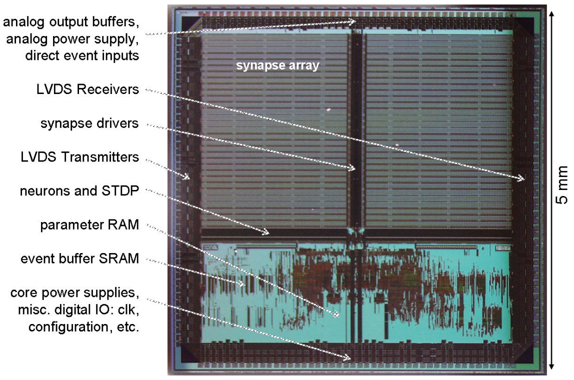

Fig. 1. Die photograph of the network chip.

one by a time-continuous recovery process. These dependencies are summarized in eq. 3:

$$\begin{aligned}\frac{dI}{dt} &= -\frac{I}{t_{rec}} + U_{\mathrm{SE}}\cdot R\cdot\delta(t-t_{AP})\\ R &= 1-I \quad , \quad g_{\max} = A_{\mathrm{SE}}\cdot R\end{aligned} \tag{3}$$

In the case of facilitation the roles of I and R are exchanged and $g_{\max}$ becomes proportional to I.

C. Long Term Synaptic Plasticity

The correlation measurement for STDP is part of every synapse. It is based on the biological mechanism as described in [5][6]. For each occurrence of a pre- or post-synaptic action potential the synapse circuit measures the time Δt that has passed since the last occurrence of the respective other action potential. The exponentially weighted time difference is called the STDP modification function F [5] and is defined as follows:

$$F(\Delta t) = \begin{cases} A_+ \exp(\frac{\Delta t}{\tau_+}) & \text{if } \Delta t < 0 \quad \text{(causal)} \\ -A_- \exp(-\frac{\Delta t}{\tau_-}) & \text{if } \Delta t > 0 \quad \text{(acausal)} \end{cases} \tag{4}$$

A look-up table is used to translate the output of the modification function into a change of the synaptic strength ω, depending on the actual value of ω. Thus, the STDP implementation is not limited to an additive or multiplicative update rule [8], instead a wide range of possible update rules can be programmed into the chip.

D. Network Model

The network model is based on the transmission of events from one source neuron to multiple destination neurons, which do not need to be located on the same die. Events are communicated digitally but continuous-time inside the chip. The on-chip network is fully connected, i.e. each neuron can be connected to any other via a synapse. The only limitation thereby is the number of synapses per neuron, which is smaller than the total number of neurons. Therefore, for each neuron a subset of the chip's neurons has to be selected as input. A synapse receiving an event converts the digital pulse into an exponential onset and offset of the synaptic open probability $p(t)$ to generate the experimentally observed time-course of the synaptic conductance [9]. Events crossing the die frontier must leave the continuous-time domain; they get a digital time-stamp marking their onset.

III. Chip Overview

Fig. 1 shows a photo of the network chip. Two synapse arrays containing 50k synapses each occupy most of the area. Fig. 2 shows the operation principle of the synapse and neuron circuits. Both synapse arrays consist of 256 rows × 192 columns below which 192 neurons are located. Each neuron contains a capacitance C_m that represents the membrane capacitance. Three different conductances model the different ion channel currents.

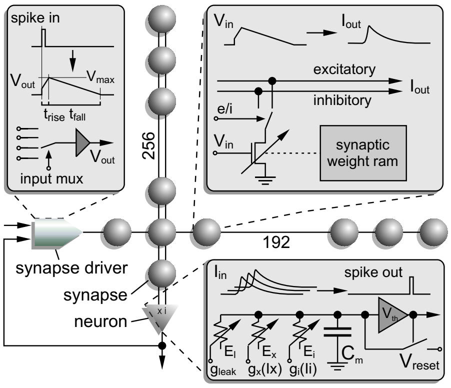

Fig. 2. Operating principle of the spiking neural network. The three boxes show the signal processing done by synapse drivers, synapses and neurons.

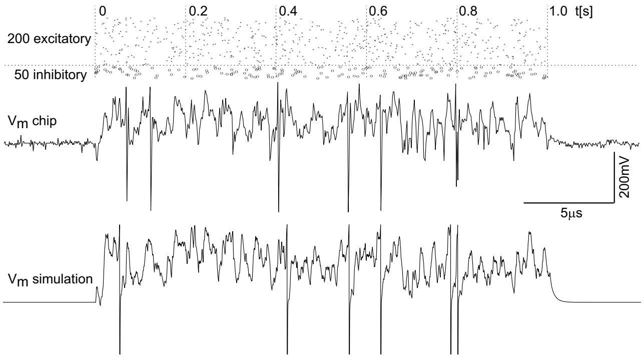

Fig. 3. Membrane potential V_m response to 250 Poisson distributed input spike trains. *Top:* Input spike trains with 3 Hz mean firing rate in biological time. *Middle:* V_m recorded from one neuron of the presented chip. *Bottom:* V_m calculated with the neuro-simulator NEST.

The membrane leakage current flows through g_{leak}. It can be individually controlled for each neuron. The leakage reversal potential E_l, the excitatory and inhibitory reversal potentials of the synapse conductances g_x and g_i as well as the threshold and reset voltages V_{th} and V_{reset} can be set for groups of 96 neurons each.

In most biological neurons, the synapse conductance is generated by the ion channels of synapses that are distributed across the dentridic tree and, to a lesser extend, the soma of the neuron. On the chip the membrane capacitance and the conductances connected to it are localized inside the neuron. The excitatory and inhibitory conductances are controlled by the sum of the currents generated by the active synapses located in the respective column. A third conductance models all ion channels contributing by their respective leakage currents to the neuron's resting potential E_l.

Whether a synapse is excitatory or inhibitory is determined by a control signal common to all synapses of one row which is used inside the synapse to switch its output between the excitatory or inhibitory input line. The weight storage of the synapse is implemented as static RAM whose content is converted into a current by a 4-bit multiplying DAC in each synapse.

Fig. 3 shows the response of a single neuron's membrane potential to Poisson distributed spike trains. The mean rate of each input was 0.12 MHz. With the onset of the input spike trains the neuron enters a high-conductance state [10] which is characterized by an increase of the membrane conductance and a depolarization of the membrane potential.

In contrast to a biological neuron the axon of its VLSI counterpart is electrically isolated from its input. It carries a digital signal that encodes the exact time of each spike's occurrence by its rising

TABLE I
CHIP SPECIFICATION SUMMARY

process features	0.18 μm, 1 poly, 6 metal
die/core size	5 × 5 mm^2/ 4.25 × 4.32 mm^2
synapse size	10.3 × 10.5 μm^2
neurons/synapses	384/98304
supply voltage (digital and analog)	1.8 V
digital core clock frequency	200 MHz
adjustable analog parameters	2969
parameter resolution	10 bit (nominal)
event time resolution (TDC, DTC)	156 ps (nominal)
event input FIFOs	16 channels, 64 entries each
event output FIFOs	6 channels, 128 entries each
LVDS bus data transfer rate	3.2 Gigabyte/s max.

edge. This signal is also routed back along the same column of synapses that comprises the neuron's input. This allows the STDP circuit located inside each synapse to measure the time between a pre-synaptic pulse and a post-synaptic spike. The digital APs are routed to the *synapse drivers* (see Fig. 2) which convert them to pre-synaptic voltage pulses controlling the synaptic conductances. The short-term plasticity circuit is also located inside the synapse driver.

Several steps are necessary to code the spikes into events. An asynchronous priority encoder identifies the spiking neuron and sends its number to the next stage. If more than one neuron fires at the same time, the neuron with the highest priority is selected. After its number has been transmitted, the one with the second highest priority gets its turn and so on. For each selected neuron, a time-to-digital converter (TDC) measures the point in time of the spike.

The digital control occupies about one-third of the core area. Its main task is to manage a set of FIFO buffers for the incoming and outgoing event signals and the formatting of event packets that can be sent and received via the LVDS external interface. Table I summarizes the specifications of the presented chip.

IV. IMPLEMENTATION OF SYNAPTIC PLASTICITY

A. Short term synaptic dynamics

Short term plasticity is realized by dynamically modifying the maximum conductance of the synapses. Fig. 4 shows the circuit diagram of the relevant part of the synapse drivers. This circuit also implements the controlled rise and fall of the synaptic conductance. Its output signal $V_{\rm out}$ directly drives the gate voltage of the current sinks of all 192 synapses located in a row (see Fig. 2). In the resting state it is pulled down to $V_{\rm rest}$ by M_7 and M_4 since the last AP left the *fall/$\overline{rise}$* signal in the high-state. Although M_9 is in a high-impedance state, the level of the *fall/$\overline{rise}$* signal is kept high by it, since M_9 is a reduced threshold-voltage device having a much larger leakage current than M_2.

An incoming pre-synaptic AP enters the circuit at the *pre* pin. It pulls the *fall/$\overline{rise}$* line down and simultaneously activates M_1. M_1 directly connects $V_{\rm out}$ to an external low-impedance voltage source $V_{\rm start}$, therefore $V_{\rm out}$ jumps from $V_{\rm rest}$ to $V_{\rm start}$. The *pre* pulse lasts only for about 5 ns. Afterwards $V_{\rm out}$ rises further at a constant rate controlled by $I_{\rm rise}$ and the current mirror built from M_5 and M_6.

The output current of the synapse changes exponentially with $V_{\rm out}$, therefore for values close to $V_{\rm rest}$, which is below 100 mV, the absolute output current of the synapse is very low. Without M_1 the linear rise of $V_{\rm out}$ would introduce a biologically unrealistic slow rise of the effective synaptic conductance. By setting $V_{\rm start}$ and $I_{\rm rise}$ accordingly, the rise time can be controlled within a very broad range. This is necessary because biological rise times may be below 1 ms for fast excitatory synapses [9].

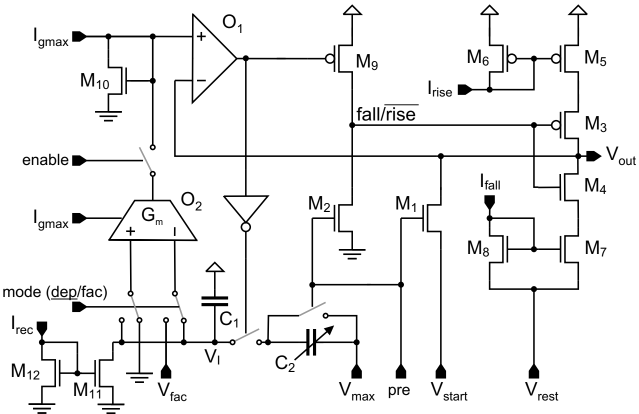

Fig. 4. Synapse driver circuits implementing short-term synaptic dynamics.

To control $g_{\rm max}$ the comparator O_1 limits the rise of $V_{\rm out}$ by pulling the *fall/$\overline{rise}$* line high. Since the real output current of the synapse can not be measured in the synapse driver, a replica transistor M_{10} is used which is an exact copy of the complete current DACs of two adjacent synapses. It is diode-connected and sinks the reference current $I_{\rm gmax}$. Its gate voltage is compared to $V_{\rm out}$ to determine the point in time when the synapse has reached its maximum output current corresponding to $g_{\rm max}$. $I_{\rm fall}$ sets the fall-time of $V_{\rm out}$ via the current mirror M_7/M_8. M_9 turns off after $V_{\rm out}$ has fallen below the gate voltage of M_{10}. The circuit is now ready for the next AP.

Synaptic plasticity is activated by the *enable* signal. The *mode* line switches between depression—as it is shown in Fig. 4—and facilitation. The output current from the OTA O_2 is added to $I_{\rm gmax}$, thus increasing or decreasing the reference voltage seen by O_1. The maximum current the OTA is able to source or sink is determined by its bias input connected to a copy of $I_{\rm gmax}$. Thus, provided the input voltage difference is large enough, the OTA will sink all the current previously sunk by M_{10}, reducing the reference voltage seen by O_1 below $V_{\rm start}$. The voltage on the node V_I represents the inactive partition I of the synaptic efficacy. After a prolonged period of inactivity V_I has dropped to zero since C_1 is continuously charged through M_{11}. With each AP charge is transferred from C_1 to C_2, increasing V_I. Therefore, from the n-th to the $(n+1)$-th AP V_I changes as follows:

$$V_{I,n+1} - V_{I,n} = \frac{C_2}{C1 + C2}(V_{max} - V_{I,n}) \qquad (5)$$

This is equivalent to the model defined by eq. 3. The capacitance of C_2 can be digitally controlled between $1/8$ and $7/8$ of C_1. The OTA translates this into a current which is subtracted from $I_{\rm gmax}$. In the case of facilitation the V_I line is connected to the non-inverting input of O_2, while the inverting input is connected to a reference voltage $V_{\rm fac}$. While eq. 5 remains valid the current subtracted from $I_{\rm gmax}$ at the input of O_1 is now proportional to $V_{\rm fac} - V_I$. Setting $V_{\rm fac}$ between $V_{\rm max}$ and zero leads to a transition from strong depression for $V_I = 0$ to strong facilitation for $V_I = V_{\rm max}$. Fig. 5 shows a simulation of V_I, the membrane potential and the output current $I_{\rm out}$ of the synapse, which is proportional to $g_{\rm max}(t)$, for both depression and facilitation. Fig. 6 shows the measured membrane potential for a similar setup. V_I and $I_{\rm out}$ are not accessible in the real chip. To set the membrane conductance to values representing the active state [11] the neuron sees a constant background of synaptic activity. The rise and fall times of the synapses participating in this background are set to

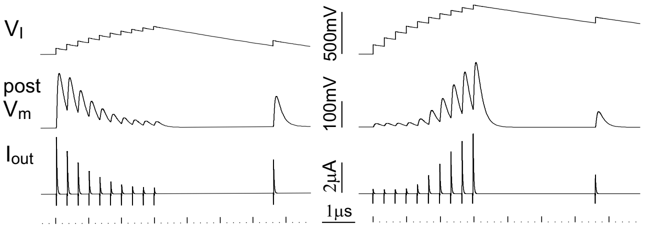

Fig. 5. Simulation results showing synaptic depression (*left*) and facilitation (*right*).

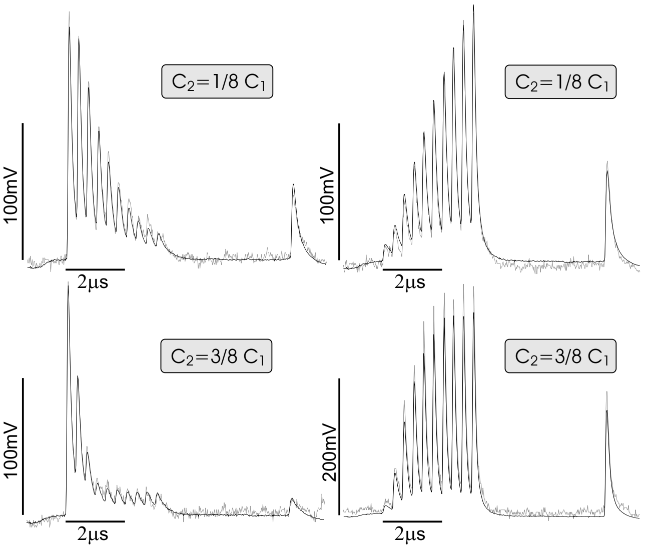

Fig. 6. Measurement of the postsynaptic membrane voltage V_m showing synaptic depression (*left*) and facilitation (*right*). The top and bottom traces differ by the C_2/C_1 ratio. The grey traces show single measurements while the solid traces are averaged results from 500 runs.

much larger values than those of the synapse under examination. Therefore their post-synaptic potentials are weak and spread over a much longer time as displayed in Fig. 6. Thus, they are not discernible in the plots of V_m. Two different values of the ratio C_2/C_1 are shown.

In the bottom case it can be seen that a steady state is reached when V_I approaches $V_{\max}$. This is caused by an equivalence between the amount of charge leaking through M_{11} between two APs and the charge transfer from C_1 to C_2 with each AP. In the case of the larger C_1/C_2 ratio this equivalence is reached after less than the 10 APs used in the experiment. By this mechanism the steady state value of $g_{\max}$ becomes dependent on the input frequency. For shorter periods between two consecutive APs less charge leaks from the *inactive* to the *active* partition, thus, the equilibrium value of V_I gets larger.

B. Spike time dependent plasticity

The synapse circuit implements eq. 4 in each synapse, thereby performing the correlation measurements fully in parallel (see [12] for details). Each synapse internally adds up these exponentially weighted measurement results independently for pre-post (causal) and post-pre (acausal) pairs. Therefore eq. 4 is broken up in two independent parts:

$$\begin{aligned} F_c(\Delta t) &= A_+ \exp(\tfrac{\Delta t}{\tau_+}), \qquad F_a = 0 \text{ if } \Delta t < 0 \text{ (causal)} \\ F_a(\Delta t) &= A_- \exp(-\tfrac{\Delta t}{\tau_-}), \quad F_c = 0 \text{ if } \Delta t > 0 \text{ (acausal)} \end{aligned} \tag{6}$$

F_c and F_a are added to the accumulated modification function ΣF_c and ΣF_a. The values of $\Sigma F_{a,c}$ are periodically read out by the digital STDP controller. It subsequently performs two comparisons:

$$\begin{aligned} &a) \quad |\Sigma F_c - \Sigma F_a| && > V_{\omega \text{ update threshold}} \\ &b) \quad \Sigma F_c - \Sigma F_a && > 0 \end{aligned} \tag{7}$$

If the outcome of (7a) is true, a weight update is performed by replacing the actual 4 bit weight ω of the synapse by the value stored in the row corresponding to said weight ω of the causal or acausal lookup table. Which table is used is determined by the comparison in (7b): if the accumulated value of the causal modification function ΣF_c is larger than the acausal value ΣF_a, the causal is used and vice versa. After the weight has been updated the accumulated values $\Sigma F_{a,c}$ are set to zero. If no weight update was performed, ΣF_c and ΣF_a would continue to sum up the results of the $F_{a,c}$ measurements.

V. Conclusion

In this paper a new VLSI model for biological neural systems was presented. It is based on a highly accelerated analog I&F network which implements plasticity within a large range of time scales. This is possible by combining the usage of capacitive storage for short-term variables, i.e. the inactive partition I or the accumulated modification function $\Sigma F_{a,c}$, with digital memory for persistent information like the synaptic weights. The high acceleration factor keeps the necessary capacitances small and strongly reduces leakage and fixed-pattern noise problems. First measurements show that the response of the chip is in good agreement with simulation as well as biological findings.

References

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

[2] R. Serrano-Gotarredona, M. Oster, P. Lichtsteiner, A. Linares-Barranco, R. Paz-Vicente, F. Gomez-Rodriguez, H. K. Riis, T. Delbrück, and S.-C. Liu, “AER building blocks for multi-layer multi-chip neuromorphic vision systems,” in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds. Cambridge, MA: MIT Press, 2006, pp. 1217–1224.

[3] P. D. and L. Abbott, *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems.* Cambridge, Massachusetts: The MIT Press, 2001.

[4] M. Tsodyks and H. Markram, “The neural code between neocortical pyramidal neurons depends on neurotransmitter release porbability,” *Proceedings of the national academy of science USA*, vol. 94, pp. 719–723, Jan. 1997.

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

[6] G. Bi and M. Poo, “Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type,” *Neural Computation*, vol. 9, pp. 503–514, 1997.

[7] H. Markram, Y. Wand, and M. Tsodyks, “Differential signaling via the same axon of neocortical pyramidal neurons,” *Proceedings of the national academy of science USA*, vol. 95, pp. 5323–5328, Apr. 1998.

[8] M. van Rossum and G. Turrigiano, “Correlation based learning from spike timing dependent plasticity,” *Neurocomputing*, vol. 38-40, pp. 409–415, 2001.

[9] A. Destexhe, Z. Mainen, and T. Sejnowski, “Synthesis of models for excitable membranes, synaptic transmission and neuromodulation using a common kinetic formalism,” *Comput Neurosci.*, vol. 1(3), pp. 195–230, 1994.

[10] A. Destexhe, M. Rudolph, and D. Pare, “The high-conductance state of neocortical neurons in vivo,” *Nature Reviews Neuroscience*, 2003.

[11] M. Shelley, D. McLaughlin, R. Shapley, and D. Wielaard, “States of high conductance in a large-scale model of the visual cortex,” *Journal of Computational Neuroscience*, vol. 13, pp. 93–109, 2002.

[12] J. Schemmel, A. Gruebl, K. Meier, and E. Mueller, “Implementing synaptic plasticity in a VLSI spiking neural network model,” in *Proceedings of the 2006 International Joint Conference on Neural Networks (IJCNN’06).* IEEE Press, 2006.