2006 International Joint Conference on Neural Networks
Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada
July 16-21, 2006

Implementing Synaptic Plasticity in a VLSI Spiking Neural Network Model

Johannes Schemmel, Andreas Grübl, Karlheinz Meier and Eilif Mueller

Abstract— **This paper describes an area-efficient mixed-signal implementation of synapse-based long term plasticity realized in a VLSI[1] model of a spiking neural network. The artificial synapses are based on an implementation of** *spike time dependent plasticity* **(STDP). In the biological specimen, STDP is a mechanism acting locally in each synapse. The presented electronic implementation succeeds in maintaining this high level of parallelism and simultaneously achieves a synapse density of more than 9k synapses per mm^2 in a 180 nm technology. This allows the construction of neural micro-circuits close to the biological specimen while maintaining a speed several orders of magnitude faster than biological real time. The large acceleration factor enhances the possibilities to investigate key aspects of plasticity, e.g. by performing extensive parameter searches.**

I. Introduction

The most common contemporary approach to the modeling of artificial spiking neural networks is the numerical simulation. An alternative is their implementation in a physical model, which leads to the concept of an analog VLSI neural network. In a physical model, important physiological quantities, like the membrane potential, should be assigned an equivalent physical quantity. Currently, VLSI is the only physical system with which it is feasible to model a neural circuit. Several such implementations of neurons and synapses have been reported [1][2]. In these cases the motivation was not primarily the speed but the continuous-time behavior. The approach presented in this paper focuses on an analog VLSI architecture as the starting point of a new kind of fast, continuous-time neural model [3] that could complement digital simulations.

It is planned to use this system within the FACETS project [4], an interdisciplinary endeavor to model cortical microcircuits from the primary visual cortex in analog VLSI as well as numerical simulations. The neural network hardware presented in this paper should be used as a prototype platform to evaluate the transferability of such simulations to analog microelectronics.

The chosen neural model allows a description of the majority of the cortical neuron types [5] neglecting their spatial structure. It is based on a capacitive membrane model with a linear correspondence between the biological and the model membrane potential. If the membrane potential reaches a threshold voltage, a spike generation process will be triggered. To reproduce near-threshold behavior observed experimentally [6], an integrate-and-fire model is not adequate. Therefore the neuron circuit was designed in a way that it depends not only on the membrane voltage, but on its derivative as well. There is one important deviation from the biological example in the microelectronic model: speed. All time-constants are reduced by a factor of 10^5 in the presented chip, i.e. 10 ns in the model are equal to 1 ms biological time.

A biologically plausible network model must take into account the strong variations of the individual neuron's properties [7][8]. In the presented chip this is done by storing about 3000 different analog parameters in a dedicated on-chip memory. An integrated digital-to-analog converter in conjunction with a network of analog current and voltage memories distributes these signals to their target neurons. The continuous update of said analog memories is performed by a dedicated control circuit which does not interfere with the operation of the network itself. Thus the neurons' parameters can be changed during normal network activity.

Unlike in biology, an action potential is not generated by membrane ion channels but by an electronic circuit monitoring the membrane potential. To facilitate the communication between the neurons, the action potential 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. To reproduce the time course of the synaptic conductance, it is modulated by an exponential onset and decay. Similar to in-vitro and in-vivo measurements [9][10], the shortening of the membrane time constant when the total synaptic conductance reaches the high-conductance region could be studied with this model.

One plasticity mechanism which has received much attention in recent years is *spike time dependent plasticity* (STDP) [11][12]. In this model each synapse measures the temporal correlation between pre- and post-synaptic action potentials (aka. spikes) which is then used to calculate long term changes in the synaptic weights. While STDP is implemented on the circuit level in the presented system, slower adaptation processes—like neuro-modulators—as well as developmental changes in neuronal connectivity will be integrated in the digital control of the analog continuous-time model. On the timescale of less than a second, measured in biological real time, the synapse model includes short term facilitation and depression [5], which may help to stabilize the operating point of the network.

The digital nature of the spike allows the usage of standard communication protocols to interconnect individual microchips to large neural systems and to interface them with

Johannes Schemmel, Andreas Grübl, Karlheinz Meier and Eilif Mueller are with the Kirchhoff Institute for Physics, University of Heidelberg, Im Neuenheimer Feld 227, 69120 Heidelberg, Germany (email: schemmel@kip.uni-heidelberg.de)

[1]very large scale integration

0-7803-9490-9/06/$20.00/©2006 IEEE

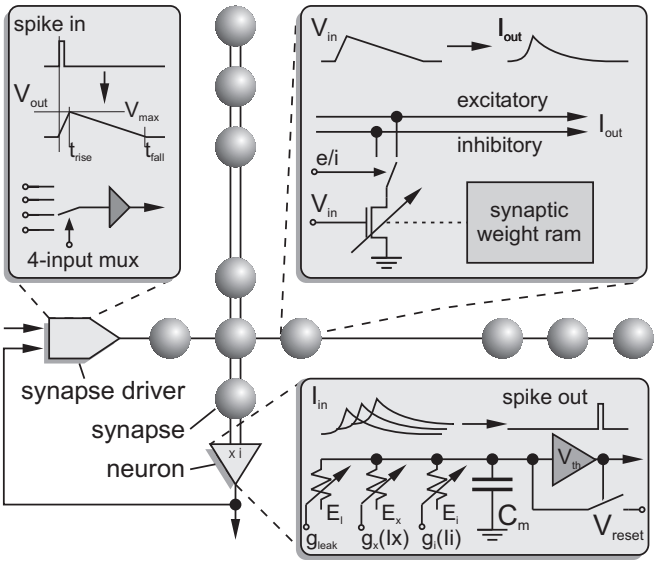

Fig. 1. Operating principle of the spiking neural network. The three boxes show the signal processing done by synapse drivers, synapses and neurons.

a digital control system to feed in data and to extract results. The presented chip uses two unidirectional 16 bit wide LVDS links with a data transfer rate of 1.3 GBytes/s each to send and receive digitized action potentials, subsequently called events.

II. Utilized Models

A. Neuron Model

The membrane potential V is governed by the following differential equation:

$$c_m \frac{dV}{dt} = g_m(V - E_l) \quad + \sum_k p_k(t) g_k (V - E_x) + \sum_l p_l(t) g_l (V - E_i) \qquad (1)$$

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. 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 parameters $p_{k,l}(t)$. Plasticity is included in the model by varying g_k and g_l slowly with time. The synaptic weight $\omega_{k,l}(t)$ denotes the relative synaptic strength at a given time t:

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

B. Network Model

The network model is based on the transmission of events from one source neuron to multiple destination neurons, which need not be located on the same die. Events are communicated digitally but continuous-time inside the chip. To connect an event with the neuron activation parameter $p(t)$, a special circuit is used that approximates the time course of the synaptic conductance by generating an exponential onset and offset each time an event arrives at the synapse.

Events crossing the die frontier must use some kind of transport protocol. A widely used protocol for VLSI neural networks is the *Address Event Representation* (AER) [13][14], which makes use of the fact that electronic communication is several orders of magnitude faster than the firing rates of biological neurons. AER maintains the asynchronous nature of the neural action potential. Unfortunately, AER is not suitable for the presented chip. The mean firing rate of a neuron in the chip is about 10 MHz due to the scaled neuron time-constants. This significantly reduces the speed advantage of electronic communication that makes AER attractive.

Since AER transports neural events in continuous time, all phase errors accumulated by the signal while traveling across an AER link, manifest themselves as temporal uncertainties of the inter-neuron connections in the modeled network. To keep these uncertainties below 0.1 ms (in biological time), the total phase error must be smaller than 1 ns. This is an unfeasible goal to achieve in a large and distributed system, interconnecting ten or more network chips. Therefore the neural events leave the continuous-time domain when crossing the die frontier; instead they get a digital time-stamp marking their onset.

For the subsequent transport of these digitized events from the network die to their final destination, the maximum latency is the only condition that must be met by the transport network. Since the time-stamp resolution is 150 ps, a temporal precision of $15\mu s$ (in biological time) can be maintained. Any transmitted event must arrive at the target network chip within the time interval the respective action potential would have traveled on a real axon, taking into account the acceleration factor between the micro-electronic neuron and its biological counterpart. With a factor of 10^5 (see next section) the resulting maximum transport latency is in the range of 50 to 500 ns, depending on the length of the modeled connection.

III. Chip Overview

A hardware model that intends to complement digital simulations should make good use of the available resources like silicon area and electrical power. The majority of the implemented devices is close to minimum-size geometry which implies low parasitic capacitances. The bias currents are set in a way to keep the transistors out of sub-threshold operation—to reduce fixed-pattern noise—while limiting the total analog power consumption below the point of self-heating. These constraints automatically lead to a membrane time constant much smaller than in biology. In the presented VLSI model, the scaling factor for time is 10^5, i.e. 10 nanoseconds model time correspond to one millisecond real-time. The limiting factor is the communication bandwidth

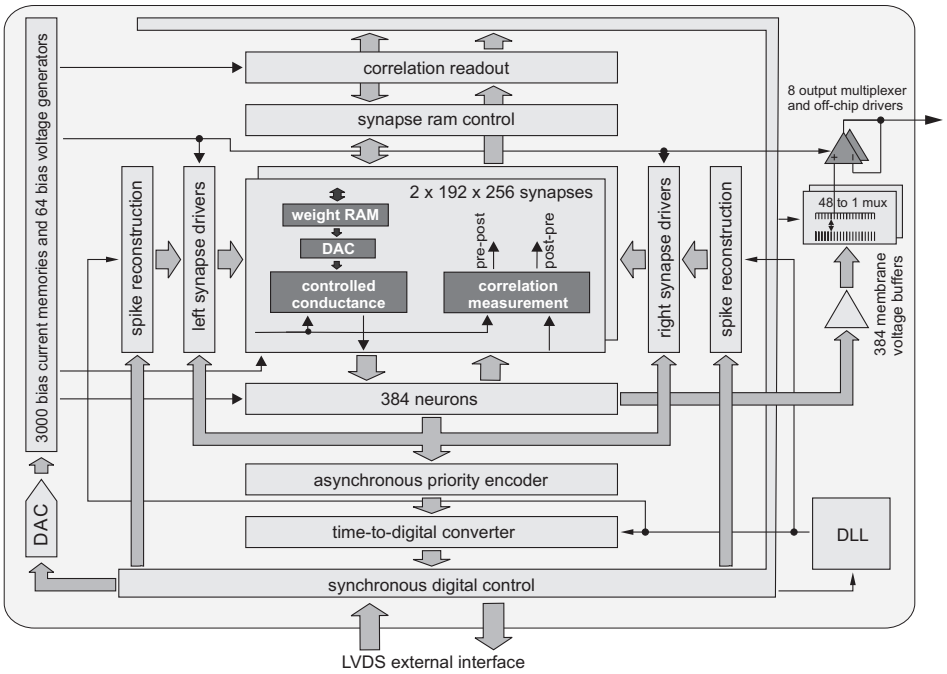

Fig. 2. Block diagram of the presented neural network chip.

since a certain number of the generated spikes must be communicated off-chip.

The weight storage of the synapse is implemented as static RAM. Compared to a capacitance-based solution [15][16] this requires a larger silicon area but facilitates a continuous-time operation spanning more than a few milliseconds. Additionally, for the built-in STDP—where the current synaptic strengths have to be modified locally—it is not possible to refresh them from an external memory which can not reflect the adapted internal values. In contrast, a digital memory placed within each synapse can be updated without having to transmit the modified weight off-chip.

Fig. 1 shows the operation principle of the synapse and neuron circuits. The STDP circuits will be described in Sec. IV and are therefore omitted here. The synapses form an array of 256 rows $\times$ 384 columns below which 384 neurons are located. Each neuron contains a capacitance C_m that represents the membrane capacitance. Three different conductances model the different ion channel currents. 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 64 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. Since no area and speed efficient solution is reported so far that mimics this organization, a different model was developed for the presented chip. Each neuron receives its input signals from one column of the synapse array. Two separate conductances are connected to the membrane capacitance inside the neuron circuit, one representing the excitatory and one the inhibitory synapses' ion channels. Each is controlled by the sum of the currents generated by the active synapses located in the respective synapse column. A third conductance models all ion channels contributing by their respective leakage currents to the neuron's resting potential E_l.

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 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.

A block diagram of the complete chip is depicted in Fig. 2. The synapse array is organized in two independent blocks of 192 columns each. This allows to place the synapse drivers in the middle column of the chip which reduces the wiring capacitances for the incoming event signals.

To code the spikes into events several steps are necessary. 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, the *time to digital* (TDC) converter measures the point in time of the spike.

A die photograph of the fabricated chip is depicted in Fig. 3. The digital control occupies about one-third of the core area. Its main task is to manage a set of FIFO buffers for

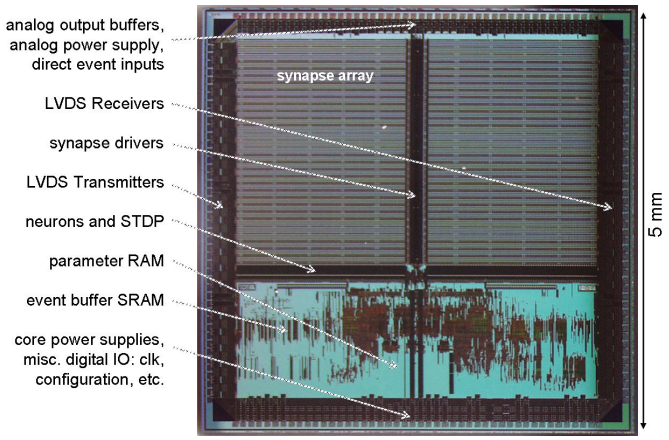

Fig. 3. Die-photograph of the presented neural network chip.

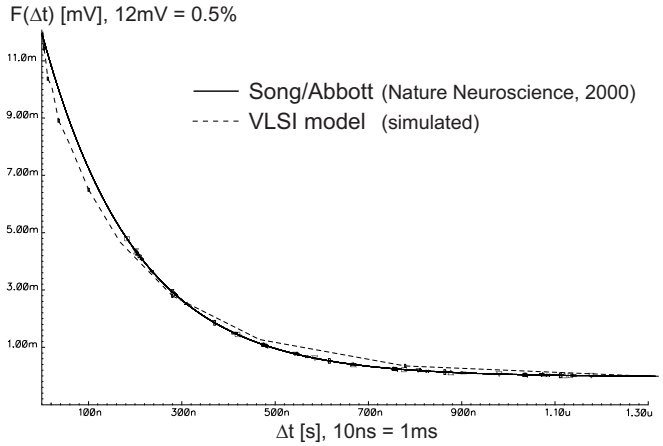

Fig. 4. Comparison between the measured modification function from [11] and the presented VLSI model.

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	2.6 Gigabyte/s (effective)

the incoming and outgoing event signals and the formatting of event packets that can be sent and received via the LVDS external interface. The LVDS links are placed along the left and right edges of the die in a way to facilitate the interconnection of multiple network chips in a daisy-chain-like fashion. It should be noted that the neurons themselves use only a very tiny fraction of the total die area, which is dominated by the synapse circuits. Table I summarizes the specifications of the presented chip.

IV. IMPLEMENTATION OF STDP

The correlation measurement for STDP is part of every synapse. It is based on the biological mechanism as described in [11][12]. For each occurrence of a pre- or post-synaptic action potential the synapse circuit must change the synaptic strength by a factor of $1 + F(\Delta t)$. F is called the STDP modification function [11] and is defined as follows:

$$F(\Delta t) = \begin{cases} A_+ \exp(\frac{\Delta t}{\tau_+}) & \text{if } \Delta t < 0 \quad \text{(causal)} \\ -A_- \exp(-\frac{\Delta t}{\tau_-}) & \text{if } \Delta t > 0 \quad \text{(acausal)} \end{cases} \tag{3}$$

Experimental data suggest a value of about 20 ms for the time constants for causal, τ_+, and acausal, τ_-, events. The parameters A_+ and A_- have also been determined experimentally by dividing the total modification of the synaptic strength measured for multiple spike pairs by the number of pairs. Fig. 4 shows the causal modification function for $A = 0.005$. The synapse strength ω changes with each pre- or postsynaptic action potential according to eq. 4:

$$\omega_{\text{new}} = \omega_{\text{old}}(1 + F(\Delta t)) \tag{4}$$

Since STDP happens locally at every synapse without any restrictions about the number of neurons firing simultaneously, the circuit implementing eq. 4 must be located in every synapse. With the given values of $A_\pm$ the changes in the synaptic strength ω implied by eq. 4 are small compared to the maximum synaptic strength. Therefore a digital implementation of the weight storage would need at least 12 bit resolution to implement eq. 4 with adequate precision. Since this is not a feasible solution for a local per-synapse weight storage a mixed digital and analog technique is required. In the presented chip the synaptic weights are stored with four bit resolution in each synapse and the digital-to-analog converter is implemented as a digitally controlled current source using binary weighted transistors. A possible approach would have been the addition of a voltage controlled current source in parallel to the digital controlled one. Its control voltage could be stored on a capacitance and modified according to the discussed STDP modification function. Each time the current on this source would cross a threshold of one lsb above or below its starting value the digital memory would have to be incremented or decremented accordingly and the voltage controlled source be reset.

This straight forward implementation poses two problems: First, the implementation of the digital increment and decrement logic takes up too much silicon area and second, both current sources and the comparator must be calibrated to avoid jumps in the weight value when the memory gets incremented or decremented. An additional difficulty arises from the fact that $F(\Delta t)$ should become $F(\Delta t, \omega)$ to allow the

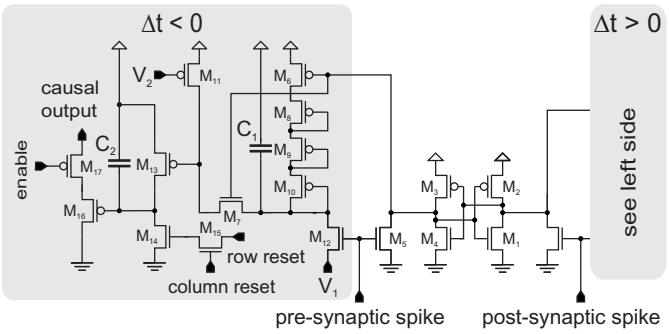

Fig. 5. Circuit diagram of the STDP circuit located in each synapse.

modeling of saturation effects if ω approaches its maximum resp. minimum value.

Therefore a special hybrid solution has been developed which is implemented in the presented chip. It uses an area efficient circuit implementing eq. 3 in each synapse, thus performing the correlation measurement fully in parallel. Eq. 4 is calculated in a mixed digital/analog circuit on the periphery of the synapse array in a sequential fashion using programmable lookup-tables to implement the weight dependence of the STDP modification function F. To avoid loosing any results each synapse accumulates the resulting F values until the sequential readout addresses the row it is located in.

Fig. 5 shows the STDP circuit which is part of the synapse. It occupies an area of only $5{\times}10\ \mu m^2$ and is symmetrical with respect to the pre- and post-synaptic inputs, therefore only the causal part is shown. The measurement circuit switches between two states—delay measurement and result accumulation—stored in the central latch build from M_1 to M_4. Considering the case that M_1 and M_3 are conducting, M_6 isolates the three series-connected transistors M_8 to M_{10} from the supply while M_7 connects C_1 to the supply via the current source M_{11}. The gate voltage of M_{13} is also kept high by M_{11}, and while both, row and column reset inputs are held inactive, C_2 maintains its charge. The voltage on C_2 represents the accumulated STDP modification result for all causal events that have occurred at the respective synapse since the last reset signal.

When a pre-synaptic spike arrives at the synapse, M_2 and M_4 will start conducting and the latch switches to the measurement state for the causal part. The spike signal is a positive pulse with a fixed duration of a few nanoseconds. Since M_7 now isolates C_1 from M_{11}, M_{12} can charge C_1 to the initial voltage V_1. The PMOS string built from M_8 to M_{10} is connected to the positive supply by M_6 during the measurement phase. Due to the sub-threshold behaviour of these devices the discharge characteristic of C_1 is a decaying exponential, thus implementing eq. 3. Any additional pre-synaptic spikes restart the measurement process by pre-charging C_1 again to V_1. With the arrival of a post-synaptic input the measurement phase is ended and the latch switches back to the accumulation phase for the causal side. Now C_1 gets discharged via M_7 by a constant current set with

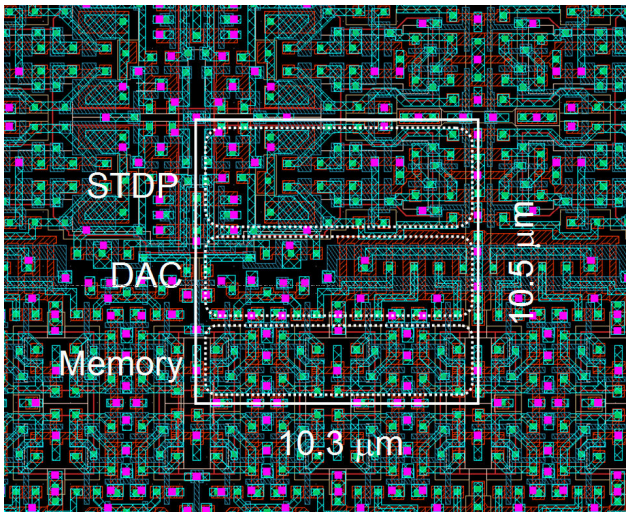

Fig. 6. Cutout from the layout drawing of the synapse array. Metal layers 2 to 6 as well as the metal-metal capacitance layer (MIMCAP) have been omitted for clarity.

M_{11}, activating M_{13} for a time span roughly proportional to the remaining charge on C_1. M_{13} in turn removes a small amount of charge from C_2 equivalent to the value of $F(\Delta t)$.

The readout of the accumulated result on C_2 is done for each synapse row in parallel via the source follower M_{16} and the row enable switch M_{17}. The STDP circuit for each synapse column compares the absolute difference between the accumulated causal and acausal values against a programmable threshold and changes the digital weight storage accordingly. In this case the row and column reset signals are activated and C_2 is charged to its reset value by M_{14} and M_{15}. Since the dynamic range of C_2 is large enough to store about 100 events there will be no loss of pre-post correlation information in normal network operation. Fig. 4 shows the result of a series of simulations from the presented circuit using different values of Δt. The obtained results and the curve that has been derived from biological data match well.

A cutout from the layout drawing is shown in Fig. 6. The solid rectangle marks the boundary of an individual synapse. Dashed boxes depict the three major functional blocks: the four bit digital weigth memory, the digital-to-analog converter and the STDP correlation measurement circuit described in the previous paragraphs. The total number of MOS transistors in a single synapse is 76. The two capacitances storing the results for the causal and acausal correlation measurements (C_2 in Fig. 5) are not visible in Fig. 6 since they are implemented as metal-metal capacitances (MIMCAP) atop of the whole synapse circuit.

Fig. 7 shows first measurements of the STDP modification function from the fabricated chip. Due to the early state of the test setup these measurements are limited to a single synapse, located at row 252 and column 66 in the left half of the chip. Pairs of pre- and post-synaptic spikes with a given time difference Δt were sent into the chip. After each pair the voltage on C_2 was compared against a threshold

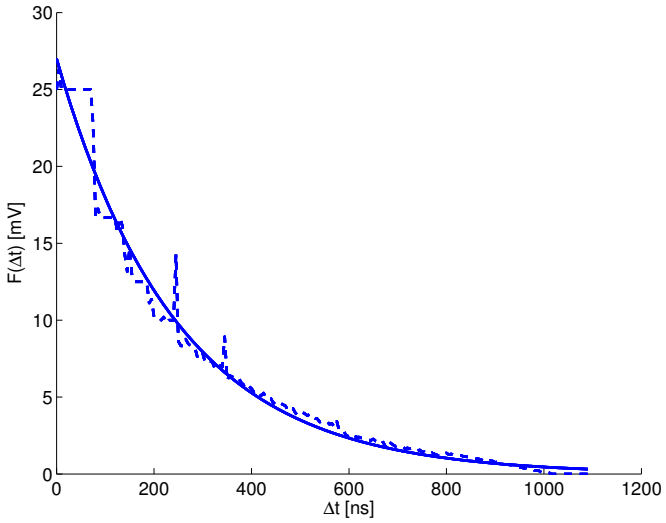

Fig. 7. Measurement of the STDP modification function for a single, arbitrary synapse (dashed line). The theoretical model is shown as a reference (solid line). The parameters were set as follows: $A = 27$ mV, $\Delta t = 245$ns.

voltage of 50 mV. The number of pairs necessary to cross this threshold was recorded. This process was repeated 10 times for each time difference Δt. After each data point Δt was incremented by 5 ns. The step-like behaviour of the activation function for small values of Δt, as it is visible in Fig. 7, is an artefact of this measurement procedure, since for small time differences only a few pulse-pairs are necessary to cross the 50 mV threshold. This results in a rather large quantization error inversely proportional to Δt. Since the threshold voltage will be increased in the final setup, this effect can be greatly reduced.

V. Conclusion

This paper presents a new approach to model synaptic long term plasticity as it is observed in biological neural systems in VLSI. By combining local temporal correlation measurement and a digitally controlled weight update, a high synapse density has been achieved while allowing adjustment of the STDP modification function using lookup tables. Since multiple chips can be connected via the 2.6 GByte/s LVDS interface, it is possible to build systems modeling in the order of 10^6 synapses. In conjunction with the high operation speed this will allow the investigation of adaptation in cortical microcircuits covering biological time spans of several minutes. Since such an experiment lasts less than a millisecond extensive parameter searches will become a new research possibility.

First measurements of the fabricated chip have been performed successfully. Especially the functionality of the STDP circuits, as described in this paper, has been verified.

Acknowledgment

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

References

[1] V. Douence, A. Laflaquiere, S. L. Masson, T. Bal, and G. L. Masson, "Analog electronic system for simulating biological neurons," in *Proceedings of the International Work-Conference on Artificial and Natural Neural Networks, IWANN 1999*, 1999, pp. 188–197.

[2] P. Häflinger, M. Mahowald, and L. Watts, "A spike based learning neuron in analog VLSI," *Advances in neural information processing systems*, vol. 9, 1996.

[3] J. Schemmel, K. Meier, and E. Mueller, "A new vlsi model of neural microcircuits including spike time dependent plasticity," in *Proceedings of the 2004 International Joint Conference on Neural Networks (IJCNN'04)*. IEEE Press, 2004, pp. 1711–1716.

[4] "FACETS - Fast Analog Computing with Emergent Transient States," www.facets-project.org.

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

[6] A. Destexhe, "Conductance-based integrate and fire models," *Neural Computation*, vol. 9, pp. 503–514, 1997.

[7] W. Maass, T. Natschläger, and H. Markram, "Real-time computing without stable states: A new framework for neural computation based on perturbations," *Neural Computation*, vol. 14, no. 11, pp. 2531–2560, 2002.

[8] A. Gupta, Y. Wang, and H. Markram, "Organizing principles for a diversity of gabaergic interneurons and synapses in the neocortex," *Science*, vol. 287, pp. 273–278, Jan. 2000.

[9] A. Destexhe, M. Rudolph, and D. Paré, "The high-conductance state of neocortical neurons in vivo," *Nature Reviews Neuroscience*, vol. 4, pp. 739–751, 2003.

[10] 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.

[11] 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.

[12] 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.

[13] M. Mahowald, *An Analog VLSI System for Stereoscopic Vision*. Kluwer, 1994.

[14] A. Mortara and E. A. Vittoz, "A communication architecture tailored for analog VLSI artificial neural networks: intrinsic performance and limitations," *IEEE Trans. on Neural Networks*, vol. 5, pp. 459–466, 1994.

[15] J. Schemmel, K. Meier, and F. Schürmann, "A VLSI implementation of an analog neural network suited for genetic algorithms," in *Proceedings of the International Conference on Evolvable Systems ICES 2001*. Springer Verlag, 2001, pp. 50–61.

[16] J. Schemmel, S. Hohmann, K. Meier, and F. Schürmann, "A mixed-mode analog neural network using current-steering synapses," *Analog Integrated Circuits and Signal Processing*, vol. 38, no. 2-3, pp. 233–244, February-March 2004.