Live Demonstration: Multiple-Timescale Plasticity in a Neuromorphic System

Christian Mayr, Johannes Partzsch, Marko Noack and René Schüffny
Endowed Chair for Highly Parallel VLSI Systems and Neuromorphic Circuits
Institute of Circuits and Systems, Technische Universität Dresden, Germany
Email: christian.mayr@tu-dresden.de, ISCAS alternative track selection: 8.1,8.8

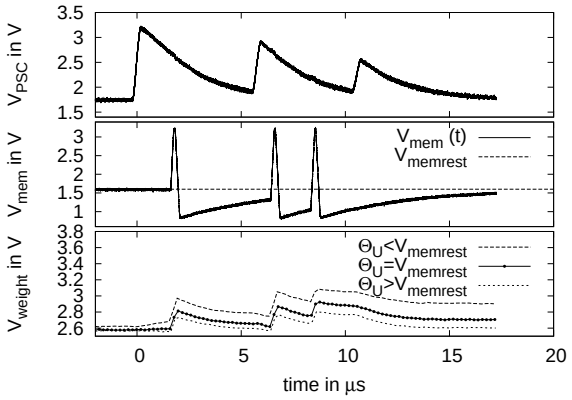

Fig. 1. From top to bottom: postsynaptic current trace (PSC) with depression, membrane voltage trace (with resting potential marked), and analog weight state progression for different settings of the learning threshold Θ_U [3]

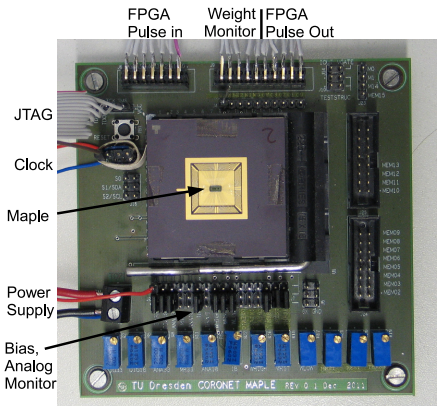

Fig. 2. PCB housing one of the Maple-ICs and support circuits. The demonstration setup links up to 15 of these boards by a standard Xilix Virtex5 experiment board.

I. Demo Description

Traditionally, neuromorphic ICs have integrated only reduced subsets of the rich repertoire of plasticity seen in biological preparations [1], [2]. The focus with respect to long term plasticity has been mostly on Spike-Time-Dependent Plasticity (STDP) [1]. Several ICs have also implemented forms of presynaptic short term dynamics, which filter synaptic pulse input, but have no influence on other timescales of plasticity. Here, we demonstrate an IC that implements short-term-, long-term-, and metaplasticity in an integrated way following [3], where these three different timescales interact to form the overall weight at the synapse. Fig. 1 shows an example presynaptic pattern with depression and the membrane trace as input for learning [3]. The resulting analog weight state shows the influence of presynaptic depression in the step increases, comparable to [1]. Also, different settings for the learning threshold exhibit a bias towards weight increase/decrease on a metaplastic (i.e. slow) timescale similar to [2]. The overall setup features several Maple-ICs of each 16 neurons and 512 of the above synapses, interlinked via FPGA-based pulse transmission. This allows network sizes of up to 200 neurons, sufficient to demonstrate the necessity for this type of learning for a range of computational neuroscience models.

II. Visitor Experience

The Multi-Maple setup is operated via a high-level Python interface, allowing the user to quickly change parameters of the learning experiments or design their own experiment. There will be a set of experiments in the code as a starting point for the user: experiments where interplay between different timescales is of essence, e.g. demonstrations of binocular rivalry [2], structural learning in recurrent networks, etc. The individually configurable timescale nature of the synapses as shown in Fig. 1 will also allow the user to experiment e.g. with different settings for the STDP windows in different parts of the network. This can lead to structural diversification/specialization [4].

Acknowledgment

This research has received funding from the European Union Seventh Framework Programme (FP7/2007- 2013) under grant agreement no. 269459 (CORONET)

References

[1] R. C. Froemke, I. A. Tsay, M. Raad, J. D. Long, and Y. Dan, "Contribution of individual spikes in burst-induced long-term synaptic modification." *J Neurophysiol*, vol. 95, no. 3, pp. 1620–1629, 2006.
[2] E. Bienenstock, L. Cooper, and P. Munro, "Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex," *Journal of Neuroscience*, vol. 2, no. 1, pp. 32–48, 1982.
[3] C. Mayr and J. Partzsch, "Rate and pulse based plasticity governed by local synaptic state variables," *Front Syn Neurosci*, vol. 2, p. 28, 2010.
[4] Y. Yamashita and J. Tani, "Emergence of functional hierarchy in a multiple timescale neural network model: A humanoid robot experiment," *PLoS Comput Biol*, vol. 4, no. 11, p. e1000220, 11 2008.

Multiple-Timescale Plasticity in a Neuromorphic System

Christian Mayr, Johannes Partzsch, Marko Noack and René Schüffny
Endowed Chair for Highly Parallel VLSI Systems and Neuromorphic Circuits
Institute of Circuits and Systems, Technische Universität Dresden, Germany
Email: christian.mayr@tu-dresden.de

***Abstract*— The information processing of neural networks depends heavily on the learning/plasticity function carried out at the individual synapses. Traditionally, neuromorphic ICs have integrated forms of Spike-Time-Dependent-Plasticity (STDP), a subset of the rich repertoire of biological plasticity. However, STDP is challenged by rate-dependent learning as well as by biological experiments exhibiting more complex timing dependency. Also, recent research shows that incorporating these biological findings in models of plastic synapses is important for various computational functions. Here, we present an IC that implements a plasticity rule based on the postsynaptic membrane potential and the profile of the presynaptic spike. This circuit moves significantly beyond the STDP in mainstream neuromorphic ICs, replicating the triplet experiments of Froemke and Dan [1] and the timing and rate experiments of Sjoestroem et al. [2].**

I. Introduction

Endowing synapses with learning ability has long been a major target for neuromorphic hardware, as this allows the IC to become plastic, i.e. adapt to changes, restructure its topology and adjust the processing function. When looking to experimental evidence, synaptic weight is changed on a medium-to-long timescale (hours-days) by spike rates [3], spike timing in various forms [1] and membrane voltage [2]. There are also various mechanisms that change synaptic weight on very short time scales (seconds, short-term-plasticity) [1] and very long time scales (days-months, metaplasticity) [4]. Only a small subset of this repertoire has been implemented in VLSI so far. Neuromorphic ICs currently focus mainly on Spike-Time-Dependent-Plasticity (STDP), i.e. the synaptic weight change caused by a pair of pre- and postsynaptic spikes [5]–[7]. Several ICs have also implemented forms of presynaptic short term dynamics [6], but these usually operate independently of the longer timescales, contrary to the interlinking seen in biology [1].

In this paper, the so-called MAPLE (Multiscale Plasticity Experiment) IC is presented. The MAPLE IC implements a combined short-term-, long-term-, and metaplasticity, where these three different timescales interact to form the overall weight at the synapse following a novel learning rule [4]. This rule can reproduce a variety of recent experimental results and is simple to implement, since it derives most of its dynamics from the neuron and the reconstruction of the incoming pulse, i.e. the postsynaptic current (PSC) profile. Thus, the synapse complexity itself can be kept to a minimum. In Sec. II-A, the overall MAPLE architecture is detailed and the basic learning rule introduced. The synaptical circuit realizing the computation of the learning function is described in Sec. II-B. Sec. III details measurement results using various biological plasticity protocols.

II. The MAPLE IC

A. System Description

The MAPLE IC implements 16 neurons with 32 synapses each, arranged in a matrix, as shown in Fig. 1A. It employs a special system architecture that separates as much functionality as possible from the individual synapse circuit, bundling it outside the synapse matrix. The individual synapse only includes those functions that depend both on the incoming (presynaptic) spikes and the neuron state. It consists of a 4bit digital weight storage with attached digital-to-analog (DA) converter and an analog computation of long-term synaptic plasticity, realising the local correlation plasticity (LCP) rule [4]. This rule combines the PSC waveform $g(t)$, the postsynaptic membrane potential $u(t)$ and a voltage threshold Θ_u in a compact rule for the change of the synaptic weight w:

$$\frac{\mathrm{d}w}{\mathrm{d}t} = B \cdot g(t) \cdot \left(u(t) - \Theta_u\right) \ , \tag{1}$$

where B is a scaling factor. This mathematical formulation is similar to the BCM rule [3]. However, instead of the rate variables in BCM, it uses local state variables directly available to the synapse. Fig. 1B shows a measured weight change for a sample spike train. At each postsynaptic spike, the weight steeply increases due to the elevated membrane potential, while it steadily decreases during the neuron's refractory period. Thus, in this simple case, weight potentiation is determined by the PSC shape and the size of the postsynaptic spike, while depression is governed by the membrane refractoriness time course and the area under the PSC curve.

The LCP rule can replicate a multitude of biological experimental results, even when combined only with simple models for PSC and membrane voltage [4]. Furthermore, it allows the direct incorporation of other adaptation mechanisms. Short-term adaptation may be expressed in the PSC and membrane voltage waveforms, while the voltage threshold may slowly change as in the BCM rule, allowing similar forms of metaplasticity.

The LCP rule facilitates an efficient hardware implementation, because it re-uses existing waveforms. This is exploited

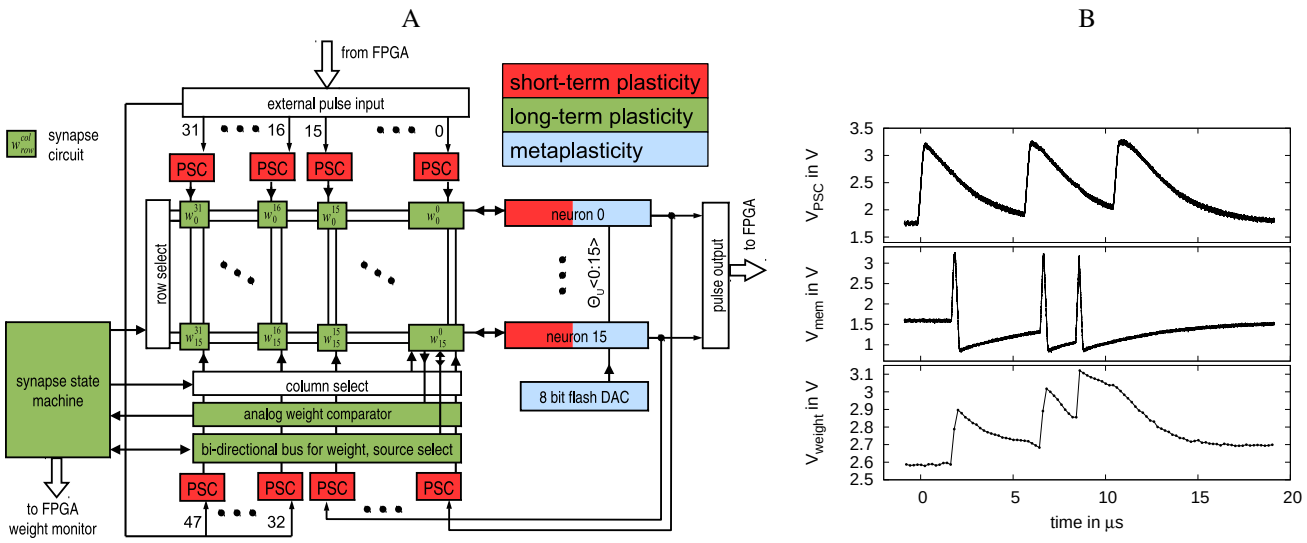

Fig. 1. (A) system architecture of the MAPLE IC, showing the different forms of plasticity implemented by the individual modules; (B) measured trace of the analog synaptic weight (V_{weight}) evolution along with the corresponding local state variables PSC (V_{PSC}) and membrane potential (V_{mem}). The time axis denotes hardware time (10^4 accelerated to biological time).

by the MAPLE system architecture described above. All functions that are only dependent on the presynaptic spikes are bundled in reconstruction circuits for the PSC, located above and below the synapse matrix. Each of them consists of a short-term plasticity (STP) circuit [8], and a generation circuit for exponentially-shaped PSCs. The number of PSC circuits is twice the number of synapse columns, so that each synapse can choose between two PSC outputs. This architecture results in a significantly more flexible synaptic connectivity at little extra cost [9].

The neuron circuit implements a leaky-integrate-and-fire model with adaptation of the postsynaptic pulse length [4], the latter directly influencing the amount of potentiation in the LCP rule. It also forwards a voltage threshold to its connected synapse row. This voltage is generated by an 8bit DA converter, so that it can be finely tuned externally during an experiment.

An external synapse state machine iterates over the synapse matrix, reading out the state of the analog weight storage and computing a new 4bit digital weight. This process is detailed in Sec. II-B. The MAPLE IC was implemented in a UMC 180nm technology, using a supply voltage of 3.3V and a typical clock frequency of 50MHz for the digital core logic. The complete chip occupies a silicon area of $1.5 \times 3.2\mathrm{mm}^2$. All circuits have been designed for a speed-up of 10^4 compared to biological time. The chip components are configured completely via a standard JTAG interface, while pulse input/output and weight output is performed via a parallel interface, allowing for simple control and monitoring by an FPGA.

B. The Plastic Synapse

Fig. 2 shows a block diagram of the synapse at the heart of the MAPLE IC as outlined in the previous section. The analog weight computation in the upper right block is detailed in Fig. 5 of [10]. It carries out the actual computation of Equ. 1 resp. Fig. 1B based on the PSC ($I_{\mathrm{psc}}, I_{\mathrm{psc\_spk}}$), the membrane voltage V_{mem} and the threshold Θ_{U}. The postsynaptic current reconstruction and its adaptation circuit is based on [8], the membrane waveform fed to the synapse is generated in a neuron based on the circuits of Fig. 2/3 of [10].

When the weight update cycle of the synaptic state machine addresses this particular synapse via the RW-Control & Select, the transmission gate M1/2 activates and the analog weight state voltage V_{weight} of this block is driven via the source follower M3 to two comparators below the synapse column. At the same time, the 4-bit digital weight (WEIGHT$<3:0>$) is read out and passed off-chip for monitoring purposes. When little weight change has happened since the last cycle, i.e. the weight voltage is between both reference voltages, nothing happens and the state machine moves to the next synapse. When the weight voltage is below $V_{\mathrm{weight\_low}}$, the state machine decrements the digital synaptic weight by an LSB, writes it in the synapse and via M4 resets the analog weight voltage to $V_{\mathrm{weight\_reset}}$ (between the two thresholds). The operation is similar for $V_{\mathrm{weight\_high}}$ and the digital weight increment. The digital weight in turn governs a current-based DA-converter. When a PSC comes in, its equivalent current timecourse $I_{\mathrm{psc\_weight}}$ is scaled with the weight value and passed as $I_{\mathrm{weight\_out}}$ to the current collection rail of the neuron of that synaptic row.

The processing chain of analog weight computation, digital readout and storage, DA-conversion and analog PSC weighting is similar to [6]. However, for the 1-bit synapse of [6], this can be done directly in the synapse. For a graded, 4-bit synapse, the digital state machine is too large to integrate into every synapse, so we adopted the approach of [7] of an external state machine.

Please note: the general makeup/functioning of the synapse

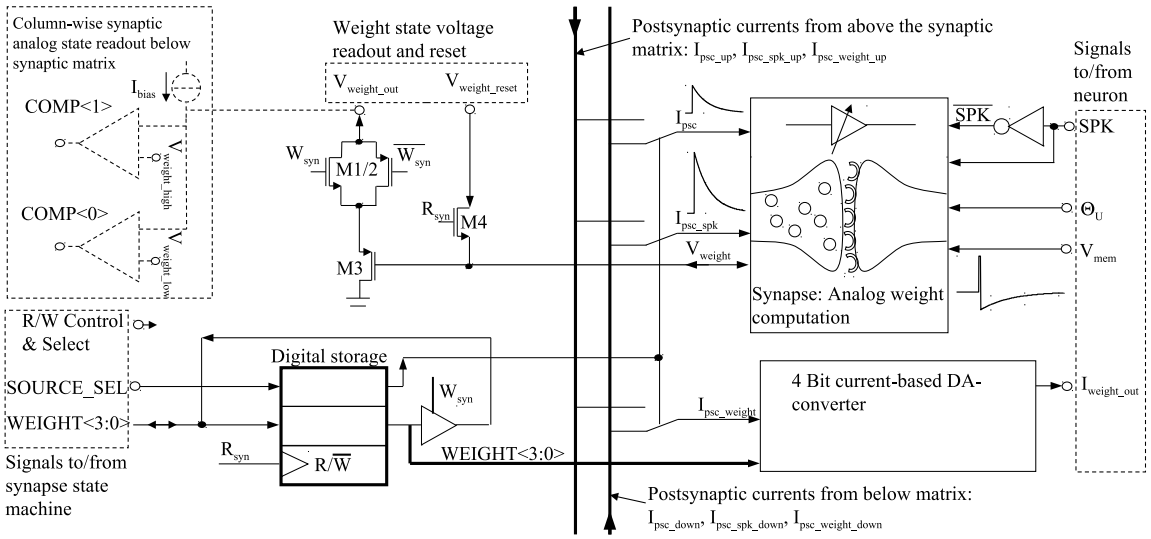

Fig. 2. Block diagram of the plastic synapse, with external ports (dashed boxes) grouped according to direction. The analog weight computation receives the digital neuron spike output, the neuron membrane voltage and PSC waveform in two different scalings. It computes the analog weight change according to the LCP rule as outlined in [10]. The upper left corner shows as an inset the immediate interface below each synapse column which converts the source-follower-driven analog weight state voltage via two thresholds to digital values. The port RW-Control & Select stands for a couple of signals that select the particular synapse for weight update and govern the read/write cycle. A small piece of logic in the synapse (not shown) converts those signals into the synapse-internal signals R_{syn} and W_{syn}. The whole synapse occupies 20*35μm^2. The digital weight bus is bidirectional. SOURCE_SEL is a static configuration bit that enables the choice between upper and lower PSC circuits (Compare Fig. 1A).

as outlined above is to a large part generic to such a mixed-signal approach for the synapse. Therefore, the building blocks surrounding the actual analog weight computation have been adapted from literature or at least constitute similar solutions to similar problems. The novelty of the MAPLE IC lies in the circuit realization [10] of weight computation based on analog waveforms supplied by neuron and presynaptic circuit. As shown in the next section, this makes the plasticity exhibited by the synapses of the MAPLE IC very biologically realistic.

III. Results

This section shows the digital weight change output for a few sample biological plasticity protocols, i.e. we are not giving the analog weight state voltage change as in Fig. 1B or as in Fig. 6 of [4], but rather the resultant digital weight change. The experiments are typical biological plasticity experiments with one synapse, stimulated via one presynaptic and one postsynaptic neuron. Please note that the results shown here are converted back from the speed-up IC time to real time for easy comparison with biological measurements.

Fig. 3 replicates conventional STDP after [1]. As outlined in Sec. II-A, the time window of the LTP part is governed by the PSC curve shape, while the LTD window depends on the membrane time constant. This is evident from Fig. 3, where different settings for PSC and neuron time constants directly influence the STDP curve.

Fig. 4 shows the digital weight change for the spike timing and frequency protocol of [2]. Ordinary STDP rules (without a frequency component) perform quite poorly at this [4], even though this crossing from LTD at low frequencies to LTP for high frequencies is crucial for various biological

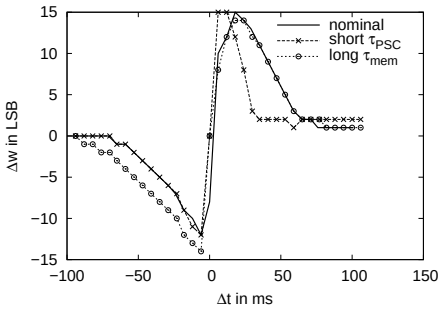

Fig. 3. Conventional STDP replication, protocol of [11]. Digital weight change curves are shown for differing settings of the PSC and membrane time constants.

functions, e.g. binocular rivalry [3]. As Fig. 4 shows, our circuit implementation results in a very good experiment reproduction of [2].

To illustrate another facet of this combined spike-timing and -rate driven plasticity, Fig. 5 gives the results for a triplet protocol [1]. To replicate the biological results in especially the lower right quadrant, a postsynaptic depression function is necessary which diminishes the influence of the second postsynaptic spike in a post-pre-post triplet, where the two postsynaptic spikes follow in quick succession (i.e. close to the origin of the diagram in Fig. 5). Fig. 5(B) illustrates that the postsynaptic depression of [10] workes as designed and, via the membrane voltage curve shape, has the same effect on the overall learning as in the ideal plasticity rule [4]. Again, the experiment reproduction is in accordance with [1], significantly better than the current state-of-the-art of other

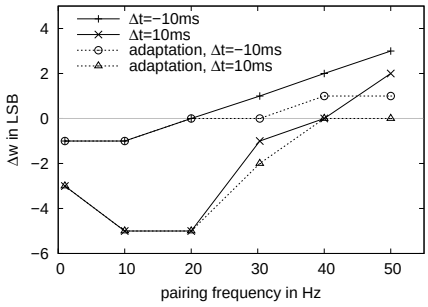

Fig. 4. Digital weight change for the combined STDP and spike frequency protocol of [2]. Note the influence of the postsynaptic adaptation on the weight change for high pulse frequencies, which is in accordance with the theoretical model [4].

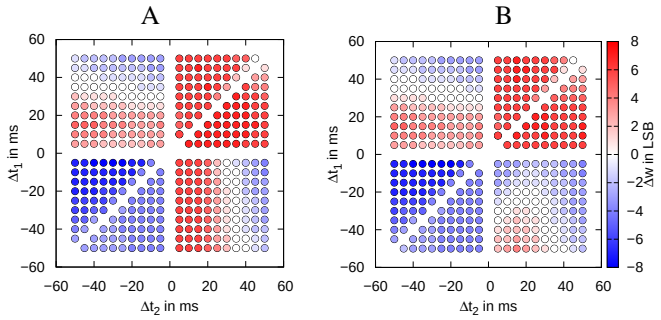

Fig. 5. (A) Digital weight change for the triplet protocol of [1] without postsynaptic adaptation of the action potential duration [10]; (B) Same as (A), with activated postsynaptic adaptation, note the difference in the lower right quadrant to (A).

VLSI-realized synaptic plasticity circuits.

IV. Conclusion

We have shown the circuit realization of a novel synaptic plasticity rule with a very high degree of biological realism. Since the learning rule is based on reusing waveform functionality which is already included in most neuromorphic systems [5], [7], circuit complexity is significantly reduced. Compared to [5], [7], [12], no time constants (i.e. capacitances) are required for the STDP time windows. Circuit area is on par with the smallest synaptic plasticity realizations currently available in CMOS [6], [7]. More importantly, even at this reduced complexity, our circuit outperforms all current synaptic learning implementations with respect to experiment reproduction, since these are only built to replicate versions of STDP, which is incompatible with spike triplet results [1] and learning evoked by combinations of timing and rate [2].

With the actual waveforms at the basis of the plasticity supplied by neuron and presynaptic reconstruction, the plasticity function of the synapse can be finely configured via e.g. the membrane time constant (see Fig. 3) or the postsynaptic adaptation (see Fig. 5). With the runtime-configurable Θ_U, we can even supply a learning bias or modulatory input to plasticity (see Sec. II-A here and also Fig. 1 in the associated demo paper), in keeping with biological evidence [3]. Beyond replicating biological experiments, this kind of multi-factor multi-timescale plasticity is important in computational neuroscience, e.g. for the learning stability or convergence of recurrent networks [13], or for developing motor primitives across multiple timescales [14].

The waveform-driven plasticity shown here also holds great promise for future nanotechnology/CMOS integration. We have recently shown that the LCP rule as implemented in CMOS circuits in this manuscript can also be applied to BiFeO3 memristive devices [15]. Together, this constitutes almost a proof of principle of a future CMOS/memristor hybrid IC, as we have shown here the veracity of the waveform generation in CMOS, and we have shown those same waveforms applied to biologically realistic learning at a memristive synapse [15]. Future work will center on actually implementing this union.

Acknowledgment

This research has received funding from the European Union Seventh Framework Programme (FP7/2007- 2013) under grant agreement no. 269459 (CORONET)

References

[1] R. Froemke and Y. Dan, "Spike-timing-dependent synaptic modification induced by natural spike trains," *Nature*, vol. 416, pp. 433–438, 2002.

[2] P. Sjöström, G. Turrigiano, and S. Nelson, "Rate, timing, and cooperativity jointly determine cortical synaptic plasticity," *Neuron*, vol. 32, pp. 1149–1164, 2001.

[3] E. Bienenstock, L. Cooper, and P. Munro, "Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex," *Journal of Neuroscience*, vol. 2, no. 1, pp. 32–48, 1982.

[4] C. Mayr and J. Partzsch, "Rate and pulse based plasticity governed by local synaptic state variables," *Front Syn Neurosci*, vol. 2, p. 28, 2010.

[5] T. Koickal, A. Hamilton, S. Tan, J. Covington, J. Gardner, and T. Pearce, "Analog VLSI circuit implementation of an adaptive neuromorphic olfaction chip," *IEEE TCAS I*, vol. 54, no. 1, pp. 60–73, 2007.

[6] S. Mitra, S. Fusi, and G. Indiveri, "Real-time classification of complex patterns using spike-based learning in neuromorphic VLSI," *IEEE Trans. on Biomedical Circuits and Systems*, vol. 32-42, no. 1, 2009.

[7] J. Schemmel, A. Grübl, K. Meier, and E. Mueller, "Implementing synaptic plasticity in a VLSI spiking neural network model," in *International Joint Conference on Neural Networks IJCNN '06.*, 2006, pp. 1–6.

[8] M. Noack, C. Mayr, J. Partzsch, and R. Schüffny, "Synapse dynamics in CMOS derived from a model of neurotransmitter release," in *20th ECCTD 2011*, 2011, pp. 197–200.

[9] M. Noack, J. Partzsch, C. Mayr, S. Henker, and R. Schuffny, "Biology-derived synaptic dynamics and optimized system architecture for neuromorphic hardware," in *MIXDES*, 2010, pp. 219–224.

[10] C. Mayr, M. Noack, J. Partzsch, and R. Schuffny, "Replicating experimental spike and rate based neural learning in CMOS," in *ISCAS*, 2010, pp. 105–108.

[11] R. C. Froemke, I. A. Tsay, M. Raad, J. D. Long, and Y. Dan, "Contribution of individual spikes in burst-induced long-term synaptic modification." *J Neurophysiol*, vol. 95, no. 3, pp. 1620–1629, 2006.

[12] K. Saeki, R. Shimizu, and Y. Sekine, "Pulse-type hardware neural network with two time windows in STDP," in *Proceedings of 15th International Conference on Neural Information Processing ICONIP08*, ser. LNCS, vol. 5507, 2008, pp. 877–884.

[13] M. A. Bourjaily and P. Miller, "Synaptic plasticity and connectivity requirements to produce stimulus-pair specific responses in recurrent networks of spiking neurons," *PLoS Comput Biol*, vol. 7, no. 2, p. e1001091, 02 2011.

[14] Y. Yamashita and J. Tani, "Emergence of functional hierarchy in a multiple timescale neural network model: A humanoid robot experiment," *PLoS Comput Biol*, vol. 4, no. 11, p. e1000220, 11 2008.

[15] C. Mayr, P. Stärke, J. Partzsch, L. Cederstroem, R. Schüffny, Y. Shuai, N. Du, and H. Schmidt, "Waveform driven plasticity in BiFeO3 memristive devices: Model and implementation," in *Neural Information Processing Systems*, 2012, accepted.