A VLSI Network of Spiking Neurons with Plastic Fully Configurable "Stop-Learning" Synapses

M. Giulioni*‡, P. Camilleri†, V. Dante*, D. Badoni‡, G. Indiveri§, J. Braun† and P. Del Giudice*

*Italian National Institute of Health, Rome, Italy and INFN-RM1, Rome, Italy. URL http://neural.iss.infn.it
†Otto-Von-Guericke University, Magdeburg, Germany
‡INFN - Roma Tor Vergata, Rome Italy
§Institute of Neuroinformatics, UNI|ETHZ, Zurich, Switzerland

***Abstract*—We describe and demonstrate a neuromorphic, analog VLSI chip (termed F-LANN) hosting 128 integrate-and-fire (IF) neurons with spike-frequency adaptation, and 16,384 plastic bistable synapses implementing a self-regulated form of Hebbian, spike-driven, stochastic plasticity. The chip is designed to offer a high degree of reconfigurability: each synapse may be individually configured at any time to be either excitatory or inhibitory and to receive either recurrent input from an on-chip neuron or AER-based input from an off-chip neuron. The initial state of each synapse can be set as potentiated or depressed, and the state of each synapse can be read and stored on a computer.**

I. Introduction

The pioneering work of C. Mead [1] introduced the term "neuromorphic" for a family of hybrid analog/digital, low-power VLSI devices designed to mimic the capabilities of biological perception and information processing. We propose a neuromorphic device that emulates biological neurons and synapses to attain "learning" ability by incorporating "Hebbian-like" mechanisms of synaptic plasticity. Whether "Hebbian" learning is based on average firing rates or on individual spikes ("spike-time-dependent plasticity" or STDP) is a matter of continuing debate and a choice that strongly influences alternative designs of neuromorphic synapse circuits. The synaptic dynamics described here are spike-driven and implements a rate-based Hebbian learning, though it is compatible with some aspects of STDP.

It has been argued [2], [3] that plausible synaptic devices may assume only a limited number of stable "states" and cannot change by arbitrary small amounts. In such conditions any deterministic learning rule (e.g., a "Hebbian" rule) has been shown to yield highly unfavorable scaling laws for memory capacity [2], [3]. The intuitive reason is that newly encoded memories rapidly erase earlier memories ("palimpsest property"). Perhaps surprisingly, far more favorable scaling laws may be attained with a stochastic learning rule, in which the Hebbian prescription renders synapses merely eligible for a state change but the probability of an actual change remains low [2], [3]. Instead of envisaging independent sources of "noise" at each individual synapse, the necessary "noise" may be provided by the irregularity of the pre- and post-synaptic spike trains (as long as the network remains in an asynchronous activity regime). In short, the theoretical analysis of learning with biologically plausible synapses appears to provide a compelling computational argument for implementing neurons as spiking elements.

Associative learning in networks of spiking IF neurons with stochastic synapses has been studied both in simulation [3], [4], [5] and in neuromorphic realizations [6], [7]. However, these first efforts were limited to artificially simplistic stimulus sets (e.g., strictly non-overlapping neural representations). To extend associative learning to more realistic stimulus sets, a further modification of the synaptic rule has been proposed, informally known as "stop-learning" [8]. In this modification, synaptic changes are additionally conditioned on average post-synaptic activity being neither too high nor too low: synapses targeting too-active neurons are not further strengthened and synapses targeting too-inactive neurons are not further weakened. This additional condition becomes crucial when partially overlapping patterns of activity are to be distinguished, as it prevents excessive potentiation of synapses in the overlapping parts, which might otherwise spoil the network's ability to distinguish these patterns. The suitability of this learning strategy was demonstrated in a Perceptron-like network for linearly separable patterns [9]. Extensions of the "stop-learning" strategy to spiking networks with recurrent connectivity are currently being pursued by several groups. We implemented a preliminary version of the 'stop-learning' synapse in a previous chip [10]; the present network implementation, besides improving on several synaptic design issues, will offer a wider range of collective dynamics through a more flexibly reconfigurable architecture. Synaptic designs inspired by the same "stop-learning" principles were proposed in [10], [11], [12], and are still an active topic of investigation.

The focus of this paper is on the reconfigurability and initialization of the 16,384 synapses as well as on the ability to read, at hardware level, the synaptic state without disrupting the internal network activity.

Section 2 provides an overview of the chip architecture and Section 3 describes synapse circuits. Section 4 illustrates results concerning the ability to configure individual synapses.

II. Chip architecture and main features

We describe a chip implementing a reconfigurable network of 128 IF neurons with spike-frequency adaptation and

978-1-4244-2182-4/08/$25.00 ©2008 IEEE.

16,384 (128×128) bistable, stochastic synapses implementing a Hebbian rule with "stop-learning" (see Fig.1). The chip, designed in standard CMOS AMS 0.35 μm technology, occupies 68.9 μm^2 of Silicon. The synaptic matrix is configurable in such a way as to support either all-to-all recurrent connectivity, exclusively external AER (Address Event Representation) connectivity, or any combination of both. In addition, the initial efficacy and the excitatory or inhibitory nature of the synapses may be set individually for each synapse. The synaptic matrix is arranged in four identical 64×64 sub-matrices. As every signal entering a sub-matrix is properly buffered, these sub-matrices could in the future serve as building blocks for considerably larger chips. The chip is compliant with the AER asynchronous communication protocol widely used in the neuromorphic engineering community. Specifically, AER-based communication is handled through the PCI-AER board [13], [14] which allows four chips to be connected together (e.g. to implement a recurrent network of 512 neurons with a uniform 25% connectivity).

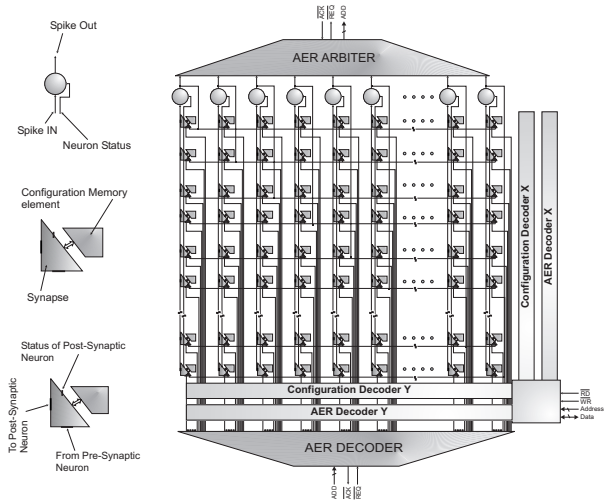

Fig. 1. F-LANN chip architecture

The neuron circuit was introduced in [15] and implements a low power IF neuron with a constant leakage term studied theoretically in [16]. An additional dynamic variable associated with the neuron reflects its recent average activity and is termed $I_{Ca}(t)$, following [8] to remind its analogy with the intracellular calcium concentration. The dendritic tree of each neuron is composed of 128 synapses. Each synapse accepts as input, spikes from either internal or external neurons. In the latter case the spikes come in the form of AER events which are addressed to the correct synapses by an XY decoder. Excitatory synapses are plastic, while inhibitory synapses have fixed weights.

Even if in principle recurrent connectivity can be achieved via the AER infrastructure, the ability to reconfigure synapses as either recurrent or AER-based allows adequate flexibility to optimally balance AER bandwidth requirements and complexity of design. A second XY-decoder allows synapses to be independently addressed and configured. In addition, the internal state of selected synapses can be set and read with dedicated circuits, independently of the AER.

The functioning of the AER has been streamlined for multi-chip systems. The AER bus is released as soon as a transparent latch array stores the AER event, without waiting for an acknowledgment from the target synapse. In this way a new transaction on the AER bus can start before the receiver chip has completely processed the previous AER event. In a multi-chip system, where spikes are delivered to different receiver chips, this mechanism increases the AER bandwidth. On the other side, if there is only one receiver chip, the AER input circuits accept a new transaction only after fixed delay, set long enough to ensure that the decoding of the previous event has been completed. On the output side of the chip, neurons reset immediately after a spike, without waiting for the Acknowledgment signal from the AER output bus. The memory of the emission of a spike is retained in a D-type flip-flop, one for each neuron, until the AER system transmits it. If the same neuron emits a second spike before the previous one has been served, than the second spike is lost. The flip-flop array decouples the internal activity from the AER one. It introduces the risk of loosing events, but ensures that AER delays do not disrupt the internal network dynamics.

III. Synapse and Calcium Circuit

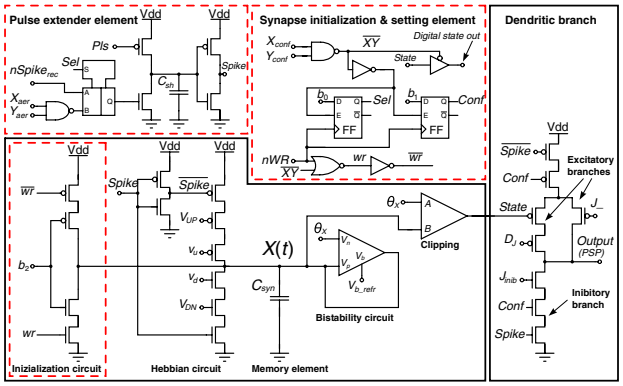

Fig. 2. Synapse circuit main blocks

Fig.2 and Fig.3 (left panel) illustrate the synaptic circuit and the comparator system needed to implement the model described in [8] and briefly motivated in the introduction. Following the arrival of a pre-synaptic spike, X jumps upward or downward, depending on the following conditions: $X(t) \rightarrow X(t) + a$ if $V(t) > \theta_p$ and $I_{TH1} < I_{Ca} < I_{TH3}$; $X(t) \rightarrow X(t) - b$ if $V(t) \leq \theta_p$ and $I_{TH1} < I_{Ca} < I_{TH2}$ where a and b are the tunable amplitudes of the jumps. In the absence of pre-synaptic spikes, if $X(t) > \theta_X$ ($X(t) < \theta_X$) X relaxes towards the upper (lower) barrier and the efficacy of the bistable synapse is set to a 'potentiated' ('depressed') state. The synaptic efficacy changes only when $X(t)$ crosses θ_X.

The *bistability circuit* (see Fig.2) is a wide output-range transconductance amplifier with positive feedback: it attracts $X(t)$ towards the upper or lower stable value depending on the comparison with the threshold θ_X, which also determines, through the *clipping-block* (a two-stage open-loop comparator), the efficacy value ($J\_$ – 'depressed' or $J\_ + DJ$ – 'potentiated'). The V_{UP} and V_{DN} signals, coming from the *calcium-block*, exclusively enable the branches of the Hebbian circuit, and inject or subtract a current regulated by v_u and v_d. The *dendritic-branch* is triggered by the pre-synaptic spike and generates the up/down jump in the post-synaptic $V(t)$ according to the configuration bit $Conf$ which sets the synapse as excitatory or inhibitory.

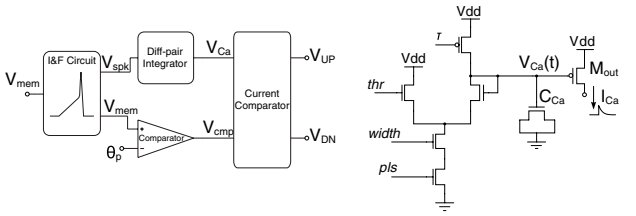

Fig. 3. Left: comparator system diagram. Right: "diff-pair integrator" log-domain calcium circuit.

The "stop-learning" mechanism relies on the "calcium" variable of the post-synaptic neuron. This variable, represented by the current $I_{Ca}(t)$, is generated by a diff-pair integrator (DPI) implementing a log-domain filter (see Fig.3, right panel). The output of this module, described in [17] and [18], is a current which increases suddenly upon the arrival of impinging spikes and exponentially decays between two spikes. For constant average firing of the neuron, the average I_{Ca} current level is proportional to the firing rate. The p-MOSFET on the right side is part of the current comparators system reported in the left panel of Fig.3. The value of I_{Ca} integrates the post-synaptic spiking activity in the recent past. Together with suitable thresholds, it determines which synaptic changes will be allowed to occur. For example, it can prevent an upward jump of $X(t)$ when the post-synaptic neuron is already very active, thus lowering the probability of synaptic potentiation.

The synapse accepts AER events (the AND of X_{AER} and Y_{AER} signals in Fig.2) or recurrent spikes $nSpike_{rec}$, depending on the configuration bit Sel. The event triggers the *pulse extender* circuit which generates a pulse $Spike$ with a tunable length. In typical conditions an AER event lasts around 200ns while the recurrent spike only 10-20ns. The circuit has the function to equalize the AER and recurrent pulses length by extending them to a few microseconds. This "long" interval of time allows, together with parameters v_u and v_d a fine tuning of the amount of charge injected or subtracted from the synaptic capacitor C_{syn}, giving rise to the jumps in X. The same interval of time determines the duration of the induced synaptic current on the post-synaptic neuron.

The module shown in Fig.3 compares I_{Ca} to three thresholds I_{TH1}, I_{TH2}, and I_{TH3} (internal to the current comparator block) to generate the two signals V_{UP} and V_{DN} shared among all synapses belonging to the same dendritic tree. The comparison is performed by three current-mode winner-take-all circuits [19], [20]. In parallel, the instantaneous voltage value of the post-synaptic neuron potential $V(t)$ is compared to a threshold θ_p (see Fig.3). Depending on the outcome of these comparisons, the current-comparator produces either an output current enabling an upward jump for $X(t)$, a current enabling a downward jump, or no output current at all. Two corresponding voltages V_{UP} and V_{DN} are produced by current-conveyors and broadcasted along the neuron's dendritic tree. This system of comparators implements the inequalities above for the dynamics of $X(t)$.

Fig.4 illustrates the effect of the calcium circuit on $X(t)$. Thresholds were set to have $I_{TH3} > I_{TH1} = I_{TH2}$. The synapse is initially set depressed and then a constant current is injected into the post-synaptic neuron. The neuron activity begins and the calcium variable I_{Ca} undergoes upward jumps moving to a new asymptotic average value. The trace labeled I corresponding to $(\frac{I_{Ca}}{I_0})^K$ where K and I_0 are two constant values, is derived from the measured V_{Ca} and computed as $I = e^{V_{dd}-V_{Ca}}$. When I_{Ca} is smaller than $I_{TH1} = I_{TH2}$, transitions of X are disallowed. In the intermediate regime between $I_{TH1} = I_{TH2}$ and I_{TH3}, up transitions are allowed. When I_{Ca} is larger than I_{TH3}, transitions of X are once again disallowed. Note that X relaxes towards its lower bound when $X < \theta_X$ and towards its upper bound when $X > \theta_X$.

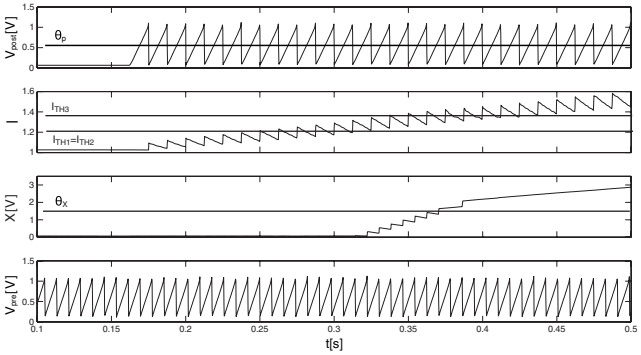

Fig. 4. Illustrative example of the 'stop-learning' mechanism (see text). Top to bottom: the post-synaptic neuron potential, the current I, the internal synaptic variable X, and the pre-synaptic neuron potential. The thresholds I_{TH1}, I_{TH2} and I_{TH3} are indicated by the dashed horizontal lines, together with the threshold θ_X.

IV. Synapse Configuration

Compared to the previous synapse designed for the C-LANN chip [10], the synaptic circuit in Fig.2 embodies new and improved blocks: the new version of the pulse extender, the improved SISE (*synapse initialization and setting element*) and the initialization circuit. This allows a higher degree of configurability and provides the ability to initialize and read the synaptic state via dedicated hardware. The SISE block, designed with AMS digital standard cells, consists of two flip-flops, used to store the configuration bits, and of a three-state buffer, necessary to read the synaptic state. A digital

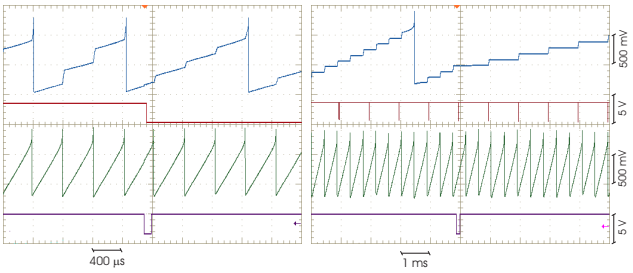

Fig. 5. Left panel: Effect of re-configuration, a potentiated synapse is set to the depressed state. Top to bottom: post-synaptic neuron potential, X signal, pre-synaptic neuron potential, nWR signal. Right panel: a recurrent synapse is set to be AER. Top to bottom: post-synaptic neuron potential, pre-synaptic AER spike, pre-synaptic neuron potential, nWR signal.

control bus carries 4 bits ($b0$, $b1$, $b2$, nWR) to the SISE selected by the row-column lines X_{conf}, Y_{conf}. The activation of the nWR bit induces both the loading of $b0$ and $b1$ in the respective FF (Flip-Flop) and the initialization of the internal synaptic variable X. Bit $b0$ loaded in the first FF produces the Sel signal which configures the synapse as either recurrent or AER, through the MUX visible in the *pulse extender element* in Fig.2. The *pulse extender element* regulates the duration of the spike; it is controlled by the voltage Pls and triggered by the incoming spike, either AER or recurrent. Bit $b1$ loaded in the second FF produces the $Conf$ signal which sets the synapse as excitatory or inhibitory through the dendritic branch. Bit $b2$, input to the *initialization circuit*, is a global signal over the entire synaptic array: it dictates the potentiated or depressed state of the selected synapse when nWR is enabled. The selection of a synapse through the X_{conf} and Y_{conf} lines enables the corresponding three-state buffer: a bit coding for the synaptic state is made available on one of the chip output pins. The digital lines involved in this signal path have been designed with particular care to reduce cross-talk effects on the slow analog lines. This allows the continuous scanning of the synaptic matrix, and hence the real time monitoring of the synaptic evolution, without disrupting the network dynamics. Such a feature greatly simplifies the experimenters' work during high-level testing. Decoders are used to access the synapses to configure them as excitatory or inhibitory, and recursive or AER. Other decoders are also used when addressing the synapses in case of AER spiking activity. 7-to-128 bit decoders were implemented to address the 128×128 synaptic matrix, using standard cells from austriamicrosystems (AMS) and automatic place-and-route tools supplied by CADENCE. These cells should lower noise and reduce ground bounce and voltage drops.

The synapses configurability is illustrated in Fig.5. Left panel shows the post-synaptic effects of a potentiated synapse being set as depressed (larger to smaller jumps induced in the post-synaptic potential). The right panel shows the effect of changing the synapse from recurrent to AER (post-synaptic jumps are first locked to the recurrently transmitted spikes, then become locked to the AER spikes). The excitatory or inhibitory nature of each synaptic contact can also be set; this feature, not shown in the figure, was already implemented in our previous chip [10].

V. Conclusions

We report here an analog VLSI building block (F-LANN) for multi-chip neuromorphic networks with a flexible architecture. Neurons and synapses feature adaptive and self-regulating properties designed for the associative learning of complex and partly correlated patterns. F-LANN, with 128 neurons and 16,384 synapses, is thought as a building block for larger networks composed of various chips linked via AER-based communication infrastructure. For this reason, the F-LANN implements an AER-compliant chip design in which each neuron can be configured to host an AER segment on its dendritic tree, which stands ready to accept spikes from external sources. To achieve maximal flexibility in setting a connection architecture, each synapse can be individually configured to be either recurrent or AER-based, either excitatory or inhibitory, and of either high or low initial efficacy. In addition, synapses may be read and set without impeding spike traffic on the AER bus. In summary, the F-LANN represents a critical step toward flexible multi-chip systems that perform associative learning of natural stimulus sets with biologically plausible components.

Acknowledgment

We thank E. Petetti for the technical support. We acknowledge the EU support through the IST2001-38099 grant.

References

[1] C. Mead, *Analog VLSI and neural systems*. Addison-Wesley, 1989.
[2] D. Amit and S. Fusi, *Neural Computation*, vol. 6, p. 957, 1994.
[3] S. Fusi, *Biological Cybernetics*, vol. 87, p. 459, 2002.
[4] P. D. Giudice, S. Fusi, and M. Mattia, *Journal of Physiology Paris*, vol. 97, p. 659, 2003.
[5] D. Amit and G. Mongillo, *Neural Computation*, vol. 15, p. 565, 2003.
[6] S. Fusi, M. Annunziato, D. Badoni, A. Salamon, and D. Amit, *Neural Computation*, vol. 12, p. 2227, 2000.
[7] E. Chicca and al., *IEEE Transactions on Neural Networks*, vol. 14(5), p. 1297, 2003.
[8] J. Brader, W. Senn, and S. Fusi, *Neural Computation (in press)*, 2007.
[9] W. Senn and S. Fusi, *Neural Computation*, vol. 17, p. 2106, 2005.
[10] D. Badoni, M. Giulioni, V. Dante, and P. D. Giudice, in *Proc. IEEE International Symposium on Circuits and Systems ISCAS06*, 2006, pp. 1227–1230.
[11] S. Mitra, S. Fusi, and G. Indiveri, in *Proc. IEEE International Symposium on Circuits and Systems ISCAS06*, 2006, pp. 2777–2780.
[12] G. Indiveri and S. Fusi, in *Proc. IEEE International Symposium on Circuits and Systems, ISCAS 2007*, 2007, pp. 3371–3374.
[13] V. Dante, P. D. Giudice, and A. M. Whatley, "The neuromorphic engineer newsletter," 2005.
[14] E. Chicca, V. Dante, A. Whatley, P. Lichtsteiner, T. Delbruck, G. Indiveri, P. D. Giudice, and R. J. Douglas, *IEEE Transactions on Circuits and Systems I*, vol. 54(5), pp. 981–993, 2007.
[15] G. Indiveri, E. Chicca, and R. Douglas, *IEEE Trans. on Neural Networks*, p. 211, 2006.
[16] S. Fusi and M. Mattia, *Neural Computation*, vol. 11, p. 633, 1999.
[17] C. Bartolozzi and G. Indiveri, *Neural Computation*, vol. 19(10), p. 2581, 2007.
[18] C. Bartolozzi and G. indiveri, in *Proc. Brain Inspired Cognitive Systems*, 2006.
[19] G. Indiveri, *Analog Integrated Circuits and Signal Processing*, vol. 28(3), p. 279, 2001.
[20] J. Lazzaro, S. Ryckebusch, M. Mahowald, and C. Mead, *Advances in neural information processing systems*, vol. 2, p. 703, 1989.