Ultra Low Leakage Synaptic Scaling Circuits for Implementing Homeostatic Plasticity in Neuromorphic Architectures

Giovanni Rovere*†, Qiao Ning*, Chiara Bartolozzi†, and Giacomo Indiveri*
* Institute of Neuroinformatics, University of Zürich and ETH Zürich, Switzerland
†iCub Facility, Istituto Italiano di Tecnologia, Italy

***Abstract*—Homeostatic plasticity is a property of biological neural circuits that stabilizes their neuronal firing rates in face of input changes or environmental variations. Synaptic scaling is a particular homeostatic mechanism that acts at the level of the single neuron over long time scales, by changing the gain of all its afferent synapses to maintain the neuron's mean firing within proper operating bounds.**

In this paper we present ultra low leakage analog circuits that allow the integration of compact integrated filters in multi-neuron chips, able to achieve time constants of the order of hundreds of seconds, and describe automatic gain control circuits that when interfaced to neuromorphic neuron and synapse circuits implement faithful models of biologically realistic synaptic scaling mechanisms. We present simulation results of the low leakage circuits and describe the control circuits that have been designed for a neuromorphic multi-neuron chip, fabricated using a standard 180 nm CMOS process.

I. Introduction

An increasing number of both academic and industrial research groups started investigating the design of parallel spiking neural networks VLSI devices and architectures [1], [2]. While it is known how to configure neural networks to implement arbitrary computations, it is still not clear how to build hardware neural systems that can match biological nervous systems when it comes to computing reliably and robustly, while adapting to changes in the inputs, in the environment, and in internal state variables (e.g., due to temperature drifts, network re-configuration, or occurrence of faults). Biological nervous systems carry out reliable and robust computation using massively parallel, slow, and extremely compact unreliable, analog, inhomogeneous computing elements. One of the key mechanisms that provides these systems with these remarkable features is *plasticity*. Multiple forms of plasticity have been observed in the nervous system, but only some of them have been emulated in neuromorphic VLSI technology. The most common are short- and long-term plasticity neuromorphic circuits. Such types of circuits have been used to implement mechanisms that compensate for device mismatch effects [3], [4], and for solving learning and classification tasks [1]. Another form of plasticity that can be extremely useful in engineered systems is homeostatic plasticity. In neuromorphic spiking neural networks, homeostatic plasticity can be used for compensating both temporal and spatial changes in the transistor properties (e.g., induced by mismatch effects or temperature changes), as well as for implementing more robust and powerful models of neural computation [5]. While in biology short- and long-term plasticity mechanisms operate on time scales that range from fractions of milliseconds to hundreds of milliseconds homeostatic plasticity operates on much longer time scales, ranging from seconds, to hours, or more. In this paper we propose ultra low leakage circuits and architectures that can implement long time constants to model a form of homeostatic plasticity mechanism known as synaptic scaling [6]. Homeostatic synaptic scaling is a negative-feedback, stabilizing mechanism observed in real neural circuits whereby the strength of all synapses impinging on a neuron is scaled uniformly, in order to maintain the neuron's overall spiking activity within given boundaries. This is a multiplicative effect that preserves the different ratios of synaptic weights among potentiated and depressed synapses without disrupting the effect of activity dependent learning, e.g., via Hebbian or Spike-Timing Dependent Plasticity mechanisms [7]. Due to their very long-time constant requirements, previously proposed homeostatic plasticity solutions either used floating gate devices [8] or off-chip control methods (e.g., via conventional workstations or processors) [9]. The VLSI model of synaptic scaling that we propose is based on standard and compact CMOS circuits all integrated in the same chip. Our solution is based on two main factors: (1) it uses an ultra low leakage circuit for implementing a very slow Automatic Gain Control (AGC) negative feedback loop, and (2) it exploits the properties of a Differential-Pair Integrator synapse [10] for global scaling of the synaptic weights of all synapses afferent to the silicon neuron circuit. In the next Section we present the ultra low-power AGC architecture and describe its principle of operation. In Section III we describe the details of the VLSI implementation and in Section IV we demonstrate its desired properties via circuit simulations. The full architecture has been designed and fabricated using a standard 180 nm CMOS process.

II. Synaptic Scaling as Automatic Gain Control

Typical neuromorphic computing neural architectures comprise arrays of silicon neurons, all connected to a large number of synapses. Stabilizing negative feedback mechanisms can be useful to maintain the neuron circuits in their proper operating range, in face of unwanted changes, such as temperature drifts or occurrence of faults. But these mechanisms should

978-1-4799-3432-4/14/$31.00

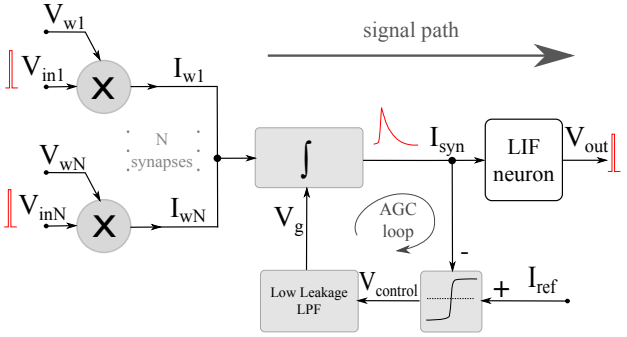

Fig. 1: Block diagram of the AGC loop. All synaptic currents sourced into the neuron are scaled by a factor that maintains the total input current to the neuron I_{syn} close to a reference current I_{ref}.

not interfere with user induced or desired changes in the neural network (e.g., due to learning or user-induced changes in parameter settings). Therefore the negative feedback loop should operate on time scales that are much larger than the ones used for parameter adaptation and learning. Furthermore, as the homeostatic synaptic scaling mechanisms acts by changing the gain of all synapses afferent to a neuron, this cannot be implemented with a classical negative feedback (additive) control loop, but requires an automatic gain control mechanism.

A. The Homeostatic Automatic Gain Control Loop

Let's assume that all synapse circuits afferent to a neuron produce an Excitatory Post-Synaptic Current (EPSC) I_{wi} weighted by their individual synaptic weight (e.g., set by a voltage bias V_{wi}). If all synapses have the same temporal dynamics then the total synaptic current received by a neuron can be written as:

$$\tau_s \frac{d}{dt} I_{syn} + I_{syn} = \gamma \sum_{i=1}^{N} I_{wi}, \tag{1}$$

where τ_s is the synapse integration time constant, γ is a global synaptic scaling factor, and N is the total number of synapses. While the changes in each I_{wi} can be governed by different types of long-term plasticity circuits [1], the synaptic scaling automatic gain control circuit should act on the term γ. In particular, this control loop should sense the changes in I_{syn} over time scales that are much larger than τ_s, and adjust γ in a way to maintain I_{syn} close to a target reference value I_{ref}. An Automatic Gain Control (AGC) scheme that implements this mechanism is shown in Fig. 1: the net input current to the neuron I_{syn} is compared to the reference current I_{ref} and a low leakage Low-Pass Filter (LPF) block slowly changes the scaling term γ to either increase or decrease I_{syn}, depending on the outcome of the comparison.

III. The VLSI implementation

Here we describe the circuits we developed to implement the AGC loop of Fig. 1.

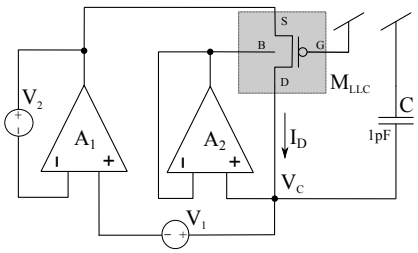

Fig. 2: Low Leakage Cell schematic.

A. pMOS Leakage Minimization

In order to obtain very long time constants in the LPF block of Fig. 1, in a compact standalone CMOS design, we minimized leakage currents across a pMOS device [11]. The drain current of a pMOS device can be described as:

$$I_D = I_{SD} + I_{BD} \tag{2}$$

where I_{SD} is the source to drain current and I_{BD} is the bulk to drain current. The term I_{SD} is strongly dependent on the pMOS operating point: if the pMOS is in accumulation mode (i.e., $V_{GB} > V_{flat-band}$), bulk majority carriers are attracted underneath the gate terminal and act as an effective insulator between source and drain terminals, thus minimizing I_{SD} leakage currents. However, in these conditions I_{SD} is still modulated by the V_{SD} potential and can vary of about one order of magnitude [12]. As a result, in accumulation mode the absolute value of I_{SD} can be minimized by setting $V_{DS} = 0$. The term I_{BD} is a reverse biased P-N junction current that is minimized by setting $V_{DB} = 0$. From these considerations, the following rules were derived [12] to obtain low leakage currents in pMOS devices as small as attoampere [12]:

$$V_{GS} > 400\,mV \quad ; \quad V_S = V_D \quad ; \quad V_{DB} = 0$$

B. The Low Leakage Cell

Fig. 2 shows a pMOS device configured with the settings described above, and with its output connected to a 1 pF state variable capacitor. The amplifiers A_1 and A_2 are low offset operational amplifiers that bias the pMOS appropriately. If $V_1 = V_2 = 0$, $V_C < V_{dd} - 400\,\text{mV}$ and the amplifiers A_1 and A_2 act as ideal unity gain buffers, then A_2 satisfies the $V_{DB} = 0$ condition and minimizes the I_{BD} current, while A_1 satisfies the $V_D = V_S$ condition and reduces the accumulation mode coupling. In this configuration, all of the conditions for a low leakage pMOS cell are simultaneously satisfied and the resulting current I_D is in the order of attoampere [12] [11]. However, while the conditions $V_{DS} = 0$ and $V_{DB} = 0$ are imposed by the circuit's topology, the condition $V_{GS} > 400\,\text{mV}$ is not always guaranteed, since $V_S = V_C$ depends on the voltage across the capacitor. This condition must be imposed by an additional circuitry that sets the DC value of V_C appropriately, to bias the pMOS device in accumulation region.

C. The Low Leakage LPF block

This block, shown in Fig. 3, comprises three sub-circuits: a voltage-offset control biasing circuit (BIAS), the low leakage

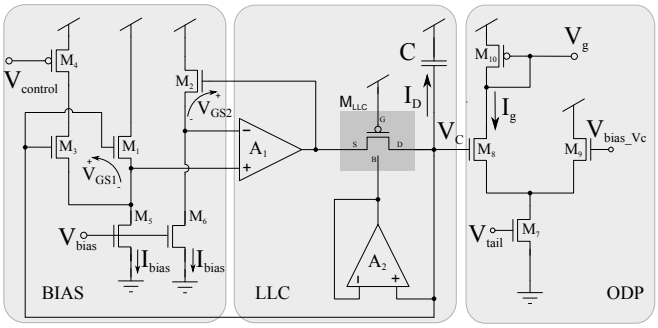

Fig. 3: The Low Leakage LPF architecture.

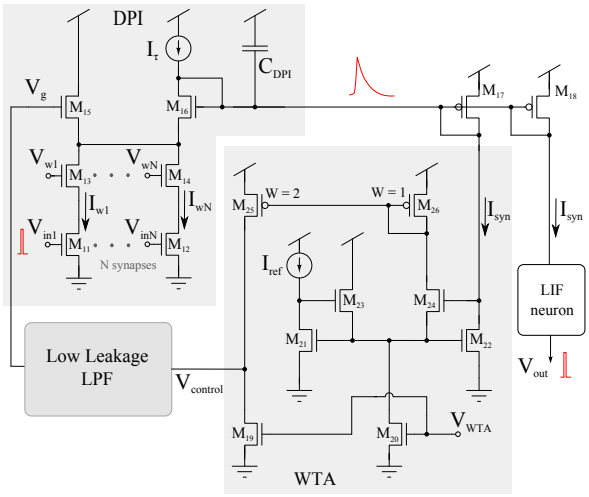

Fig. 4: Automatic Gain Control loop circuits. Synaptic input currents are integrated and scaled by the DPI block. The total synaptic current is sent to both the post-synaptic neuron and to a WTA-block that compares it to a reference current. Depending on the comparison outcome the Low Leakage LPF block slowly scales the synaptic current up or down.

cell (LLC) and an output differential pair circuit (ODP). The ODP circuit produces the output voltage V_g of the overall LPF block and once inserted in the AGC loop sets the DC voltage of V_C by biasing V_{bias_Vc} in order to satisfy the condition $V_{GS} > 400\,\text{mV}$. The LLC circuit produces the ultra low leakage current I_D as described in the previous Section. The BIAS circuit controls the offset voltage of the across A_1 amplifier in the LLC block, to set the direction of leakage current I_D, depending on the value of the $V_{control}$ signal. This signal is generated by a current-mode Winner Take All (WTA) circuit [13] that compares the total synaptic current I_{syn} to a user specified reference current I_{ref} (see WTA block of Fig. 4). The $V_{control}$ signal is high if $I_{syn} > I_{ref}$, low if $I_{ref} > I_{syn}$, and takes an intermediate value $V_{control} = V_{equilibrium}$, where $0 < V_{equilibrium} < V_{dd}$ if $I_{syn} = I_{ref}$. If $V_{control} = V_{dd}$, then the current in the M_3 branch of Fig. 3 is negligible compared to I_{bias} and results in a V_{GS1} set only by M_1 sized $(W_{eq}/L_{eq}) = (W_1/L_1)$. Conversely, if $V_{control} = 0V$, then V_{GS1} is set by an equivalent nMOS sized $(W_{eq}/L_{eq}) = (W_1/L_1) + (W_3/L_3)$. Hence, according to $V_{control}$, the equivalent size of a nMOS that provides the same V_{GS1} is indicated with (W_{eq}/L_{eq}) and takes into account the effect of the parallel between M_1 and M_3. The equation that sets V_{DS} is given by:

$$V_{DS} = V_{GS1} - V_{GS2} = -\frac{U_T}{\kappa} ln\left(\frac{W_{eq}/L_{eq}}{W_2/L_2}\right) \tag{3}$$

If the transistors M_1, M_2 and M_3 are properly sized, it is possible to set $V_{GS1} > V_{GS2}$, $V_{GS1} < V_{GS2}$ or $V_{GS1} = V_{GS2}$ according to the value of $V_{control}$. This last condition holds when $V_{control} = V_{equilibrium}$, where $0 < V_{equilibrium} < V_{dd}$ is the output of the WTA circuit when $I_{syn} = I_{ref}$.

Assuming that all transistors are matched and sized as follows: $W_1 = 1\mu m$, $W_2 = 2W_1$, $W_3 = 3W_1$ and $L = 1\mu m$, Eq.3 yields:

$$V_{DS} = \begin{cases} +V_\Delta & \text{if } V_{control} = V_{dd} \\ 0 & \text{if } V_{control} = V_{equilibrium} \\ -V_\Delta & \text{if } V_{control} = 0 \end{cases}$$

where $+V_\Delta = -\frac{U_T}{\kappa} ln(1/2)$ and $-V_\Delta = -\frac{U_T}{\kappa} ln(2)$ (which are approximately $+25\,\text{mV}$ and $-25\,\text{mV}$ respectively at room temperature).

D. The AGC loop circuits

The circuits that implement the full AGC loop are shown in Fig. 4. The total net input current received by a silicon neuron I_{syn} is compared to a reference current and, depending on the outcome, the Low Leakage LPF circuit slowly increases or decreases a gain term that is multiplied to all synaptic current. This multiplicative effect is achieved thanks to the specific type of temporal filter used to integrate the synaptic currents: a Differential-Pair Integrator (DPI) circuit [10]. In [10] the authors carry out a thorough analysis of this circuit and show that under reasonable assumptions its transfer function can be described as:

$$\tau \frac{d}{dt} I_{syn} + I_{syn} = \frac{I_g}{I_\tau} I_w \tag{4}$$

where $\tau \triangleq C_{DPI} U_T / \kappa I_\tau$ and $I_w = \sum_i I_{wi}$ represents the sum of all the weighted synaptic contributions. The current I_g does not appear in the circuit of Fig. 4, but it is the one produced by the Low-Leakage LPF circuit of Fig. 3, that sets the voltage V_g. In this context, the multiplicative term γ of eq. (1) is $\gamma = I_g/I_\tau$.

IV. Simulations Results

We simulated the AGC control circuit, together with the DPI synapse for a standard CMOS 180 nm process to validate our approach. To speed up the circuit simulations we replaced the $\sum_i I_{wi}$ input currents with a single I_{in} time varying current.

Fig. 5 shows the simulation results for an input current that abruptly changes from a steady state value of $I_{in} = 0.9\,\text{nA}$, for which the AGC loop is inactive, to a new value $I_{in} = 4.5\,\text{nA}$. In this condition the AGC loop slowly decreases I_g, to reduce the amplitude of I_{syn}, until it settles back to the value $I_{syn} = I_{ref}$. The time required to reach steady state again (Δt_{up} or Δt_{down}) lasts about 600 seconds, and is determined by the

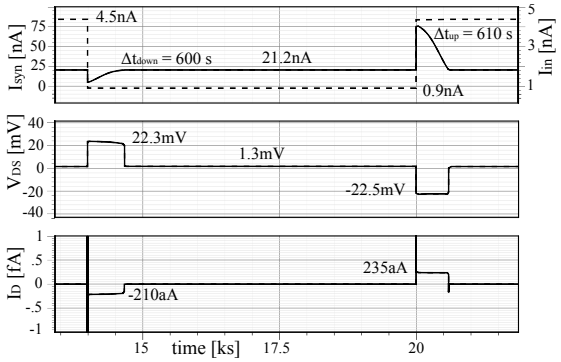

Fig. 5: Circuit simulation of slow homeostatic synaptic scaling, using a compact 1 pF capacitor. The top panel shows the DPI input and output currents. The middle panel shows the V_{DS} voltage across the M_{LLC} pMOS in the Low Leakage Cell. The bottom panel shows the M_{LLC} drain current.

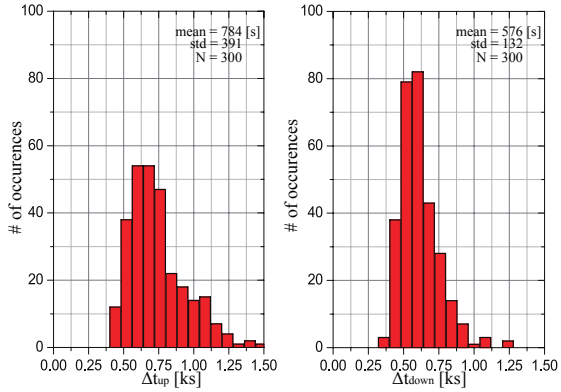

Fig. 6: The Monte Carlo simulation of transistors M_1 to M_6. The left plot is referred to an upward variation of I_{in} while the right one is referred to a downward variation of I_{in}.

current produced by the low leakage pMOS (bottom panel). As described in Section III-C, the direction and amplitude of this current is set by V_{DS} (middle panel). At steady state V_{DS} settles to $V_{DS} = 1.3mV$ to compensate for devices non-ideality and mismatches. The spikes of I_D in the bottom panel, in correspondence of steep $V_{control}$ transitions, are filtered out by the LLC capacitor of Fig. 3 and the output of the AGC is not affected by it.

In a second experiment we performed a Monte Carlo simulation to evaluate the effect of mismatch of the M_1 to M_6 MOSFETs in the I_{syn} settling time as a consequence of an abrupt change in I_{in}. The two plots of Fig. 6 shows the distributions of settling times obtained. The uneven distributions are due to the asymmetry of the $ln()$ function in eq. 3, which results in a non symmetric I_D current across the M_{LLC} pMOS, responsible for charging and discharging the capacitor C of Fig. 3.

The estimated power consumption of the AGC loop, with a supply voltage of 1.8 V, is about 100 nW.

V. Conclusions

We presented a compact low-power circuit for implementing a synaptic plasticity mechanism with very long time constants on multi-neuron chips. The long temporal dynamics required by the control loop are achieved by a low leakage cell that properly biases a pMOS. Unlike current state of the art solutions, our design does not require the use of high voltages needed for driving floating gate devices [8], nor off chip controls for obtaining ultra long time constants [9], allowing for the implementation of compact integrated solutions suitable to be used as stand alone low power device. Circuit simulations confirm the viability of our implementation, and produce responses that have settling times of over 600 seconds, far larger than the time constants used by the VLSI synapse circuits. This implementation is therefore suitable for implementing homeostatic synaptic plasticity without interfering with the network's signal processing and learning circuits.

Acknowledgments

This work was supported by the SiCode EU project, under FET-Open grant number FP7-284553, and by the neuroP EU project, under ERC grant number 257219.

References

[1] M. Giulioni, M. Pannunzi, D. Badoni, V. Dante, and P. Del Giudice, "Classification of correlated patterns with a configurable analog VLSI neural network of spiking neurons and self-regulating plastic synapses," *Neural Computation*, vol. 21, no. 11, pp. 3106–3129, 2009.

[2] T. Yu, J. Park, S. Joshi, C. Maier, and G. Cauwenberghs, "65k-neuron integrate-and-fire array transceiver with address-event reconfigurable synaptic routing," in *Biomedical Circuits and Systems Conference, (BioCAS), 2012*. IEEE, Nov. 2012, pp. 21–24.

[3] K. Cameron and A. Murray, "Minimizing the effect of process mismatch in a neuromorphic system using spike-timing-dependent adaptation," *Neural Networks, IEEE Transactions on*, vol. 19, no. 5, pp. 899–913, May 2008.

[4] J. Bill, *et al.*, "Compensating inhomogeneities of neuromorphic VLSI devices via short-term synaptic plasticity," *Frontiers in computational neuroscience*, vol. 4, 2010.

[5] A. Renart, P. Song, and X.-J. Wang, "Robust spatial working memory through homeostatic synaptic scaling in heterogeneous cortical networks," *Neuron*, vol. 38, pp. 473–485, May 2003.

[6] G. Turrigiano and S. Nelson, "Homeostatic plasticity in the developing nervous system," *Nature Reviews Neuroscience*, vol. 5, pp. 97–107, February 2004.

[7] L. Abbott and W. Gerstner, "Homeostasis and learning through spike-timing dependent plasticity," 2004.

[8] S.-C. Liu and B. Minch, "Homeostasis in a silicon integrate-and-fire neuron," in *Advances in Neural Information Processing Systems*, T. Leen, T. Dietterich, and V. Tresp, Eds., vol. 13. MIT Press, 2001.

[9] C. Bartolozzi, O. Nikolayeva, and G. Indiveri, "Implementing homeostatic plasticity in VLSI networks of spiking neurons," in *International Conference on Electronics, Circuits, and Systems, ICECS 2008*. IEEE, 2008, pp. 682–685.

[10] C. Bartolozzi and G. Indiveri, "Synaptic dynamics in analog VLSI," *Neural Computation*, vol. 19, no. 10, pp. 2581–2603, Oct 2007.

[11] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer cmos circuits." *Proceedings of the IEEE*, vol. 91, no. 2, pp. 305–327, February 2003.

[12] M. O'Halloran and R. Sarpeshkar, "An analog storage cell with 5 electron/sec leakage," in *Proc. IEEE Int. Symp. Circuits Syst*, 2006, pp. 557–560.

[13] J. Lazzaro, S. Ryckebusch, M. Mahowald, and C. Mead, "Winner-take-all networks of $O(n)$ complexity," in *Advances in neural information processing systems*, D. Touretzky, Ed., vol. 2. San Mateo - CA: Morgan Kaufmann, 1989, pp. 703–711.