# Tunable device-mismatch effects for stochastic computation in analog/digital neuromorphic computing architectures

Richard George and Giacomo Indiveri
Institute of Neuroinformatics
University of Zurich and ETH Zurich, Zurich, Switzerland
Email: [rgeorge|giacomo]@ini.uzh.ch

***Abstract*—Stochastic computing has shown promising results for low-power area-efficient hardware implementations of neural networks. In particular, probabilistic methods are being actively explored in models of spiking neural processing systems for enabling noisy and low-precision hardware neuromorphic computing architectures to implement state-of-the-art recognition and inference systems. It is therefore important to develop suitable sources of stochastic behavior for these neural processing systems that will allow them to maintain their compact and low-power benefits. Here we present a mixed-mode analog-digital circuit that can be used to control the amount of variability produced by event-based spiking neural networks, which exploits the inherent device-mismatch properties of the analog circuits used in combination with the spiking nature of the neural network. We characterize the properties of the circuit presented and demonstrate its applicability in a neuromorphic processor device comprising 256 adaptive integrate and fire neurons and 256×256 dynamic synapses.**

## I. Introduction

Neuromorphic Very Large-Scale Integration (VLSI) is a technology that aims to reproduce biophysical principles of computation that are observed in biological synapses and neurons in compact, low-power hardware [1]. One of its goals is to mimic principles of information processing in biological neural systems, such as neural computation, long-term and short-term plasticity, as well as recognition, classification, and inference. Inference neural network models can be used both for understanding the basic principles of computation used by the brain [2]–[4], and for solving practical machine learning tasks [5], [6]. However, in order to enable networks of spiking neurons to carry out probabilistic inference, it is necessary to provide a source of stochasticity that can be exploited by the neuromorphic VLSI circuits, and that does not require large amounts of silicon real-estate, or large amounts of power.

Similarly, ensemble methods in neural networks, such as “random forests” [7] or other bagging and boosting methods [8], are among the state-of-the-art methods for implementing robust classification and recognition systems. These machine learning algorithms have in common with computational neuroscience mixed-selectivity models [9] the need for a source of randomness. Also in this case, to implement such methods in neuromorphic technology efficiently (i.e., using low power and compact circuits), it is necessary to provide a source of variability that can be directly interfaced to the circuits being used to implement the neural networks. Here an ensemble of $T$ weak classifiers produce individual hypotheses $h_t$; $t \in \{1,..,T\}$ on a distribution of labeled training data $D_t$ of the overall data with an prediction error $\varepsilon_t$ with respect to $D_t$. The global classification is comprised of the weighted sum of weak hypotheses $h_t$, where the choice of $D_t$ influences the weight of the hypothesis in an adaptive fashion, detailed in [20]. The introduction of tunable spatial variability in the firing behavior of the silicon neurons proposed in this work will increase the variability in the hypotheses $h_t$, reducing the overall classification error over the training data as demonstrated in [20].

In this paper we propose a compact mixed-signal analog/digital circuit that can be used for this purpose, and that is compatible with event-based spiking neural network circuits. Although mixed-mode analog-digital, this circuit is compatible with both mixed-signal neuromorphic approaches [1], and pure digital ones [10]. The aim of this work is to investigate the effects of this circuit on the networks of spiking neurons that it is embedded in, as a step towards the exploitation and control of induced stochasticity.

In Section II we describe the principles of operation of the circuit, and the neuromorphic architecture in which we embedded it; in Section III we present experimental results to characterize its properties, and in Section IV we discuss how the results obtained can be used to implement and possibly control stochastic behavior in networks of spiking neurons.

## II. Methods

Here we first present the circuit and explain its principle of operation, and then show how it can be embedded in large-scale neuromorphic architectures to implement a tunable source of variability, potentially useful for implementing bagging or stochastic neural network models.

### A. The pulse-extender synaptic circuit

The circuit we propose is a mixed-signal pulse extender/synapse as shown in Fig. 1. It's principle of operation is based on the assumption that the efficacy of the synapse on the afferent neuron depends on both the duration of the pulse stimulating the synapse, and its synaptic weight, which sets the output current amplitude. The circuit uses two distinct

978-1-5090-6113-6/16/$31.00 ©2016 IEEE

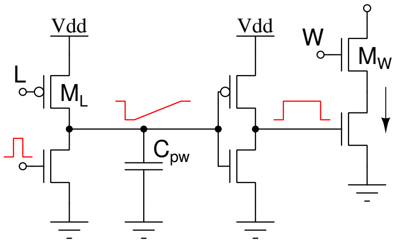

Fig. 1: Circuit diagram of a synapse pulse extender circuit. Fast input address-event pulses are fed into a starved inverter, which converts them into a waveform that is first reset and slowly recovers. The length of this recovery, and of the digital pulse generated by the subsequent inverter is controlled by the voltage Bias $L$. The output of the synapse is represented by a current $I_W$ whose amplitude depends on the Bias voltage $W$. Output currents of multiple synapses can be summed together and fed into neuromorphic circuits for implementing synapse and neural dynamics.

parameters to control these two properties: a voltage bias $L$ for controlling the duration of the pulse, and a voltage Bias $W$ for setting the weight. It is therefore possible to drive the afferent neuron is the same way by either using short pulses with large weights, or long pulses with small weights. The key intuition though lies on the fact that we can use large transistors for controlling the weight (i.e., large $M_W$) and minimum size transistors for controlling the pulse duration (i.e., small $M_L$). This will affect the amount of device mismatch present across multiple instances of these synapses in different ways: short pulses with high weights will have significantly less mismatch than long pulses with small weights.

Device mismatch is a phenomenon that affects transistors in different ways, depending on their operating domain. In particular, transistors operated in the sub-threshold domain have significantly larger mismatch than transistors operated above threshold [11], [12]. Setting a sub-threshold Bias for $M_L$ will produce mismatched currents that will integrate their differences over time to reset the input pulse, leading to a very large amount of variability in the net effect of the synapses Biased in this way. Conversely, by operating $M_L$ above threshold, (and compensating the Bias of $M_W$ to maintain the net synaptic drive $W_{eff}$), the neurons driven by these synapses will show much less variability. In Section III we quantify this variability across the neurons belonging to a Reconfigurable On-Line Learning Spiking (ROLLS) neuromorphic processor [13].

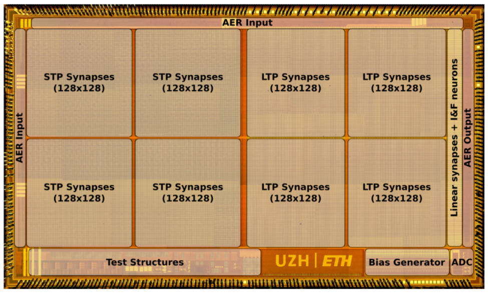

Fig. 2: Micrograph of a spike-based neuromorphic processor chip (ROLLS) that comprises the tunable mismatch pulse-with synapse circuits.

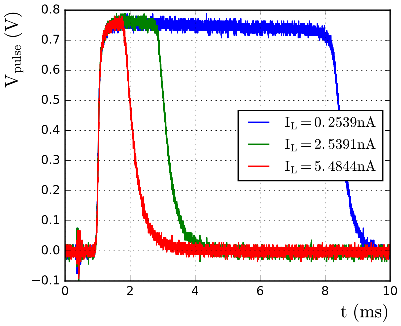

Fig. 3: Voltage output of the synapse circuit measured in response to pulses with three different lengths, for a constant weight bias setting $I_W = 0.125\,\mu A$.

### B. The ROLLS neuromorphic processor

The ROLLS chip was fabricated using a standard 6-metal 180 nm CMOS process. It occupies an area of $51.4\,\text{mm}^2$ and has approximately 12.2 million transistors. It comprises 256 neurons and 133,120 synapses. The synapse circuits are divided into arrays of short-term plasticity (STP) and long-term plasticity (LTP) elements. Both sets of synapses comprise analog circuits, that can reproduce bio-physically realistic (short-term and long-term) synaptic dynamics, as well as digital circuits that can set and change registers which control for example network configuration settings or programmable weights (see [13] for a through description and characterization of these circuits). The silicon neuron circuits implement a model of the adaptive exponential integrate-and-fire neuron [14] that has been shown to be able to accurately reproduce electrophysiological recordings of real neurons [15], [16]. All synaptic currents are integrated by low-power log-domain pulse integrator filters [17] that can reproduce synaptic dynamics with time constants that can range from fractions of micro-seconds to hundreds of milliseconds. All analog parameters of synapses and neurons can be configured via a temperature compensated programmable current-mode bias generator [18]. Peripheral asynchronous IO logic circuits are used for receiving input spikes and transmitting output spikes, using the Address-Event-Representation (AER) communication protocol [19].

Stimulation via the AER protocol is performed automatically using an FPGA device, following a predefined protocol that specifies type, target location and frequency of the input stimulus. In order to evaluate the precise timing of the neuron spikes we time-stamped the spikes as they were being measured by the FPGA device, using a timer of 40 ns resolution, and transmitted address and time-stamp of the event measured to a host PC for data logging via a USB link.

## III. Results

To determine to what extent it is possible to modulate the effect of device mismatch in the ROLLS neuromorphic processor, we stimulated one synapse per neuron and measured

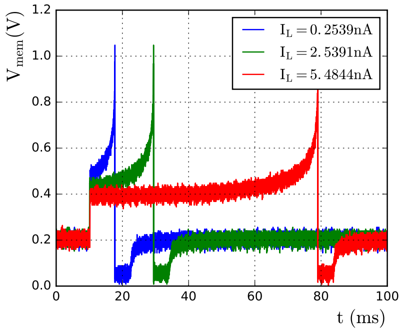

Fig. 4: Silicon neuron membrane potential measured in response to the three pulses plotted in Fig. 3, and for a constant weight bias setting $I_W = 0.125\,\mu A$.

the variability of the response across different neurons. As mentioned in Section II-B, the current produced by the mixed-mode pulse extender synapse is integrated by a first-order current mode linear filter. The final synaptic current sent to the neuron can be described by:

$$\tau \frac{d}{dt} I_{syn} + I_{syn} = \alpha I_W \tag{1}$$

where $\tau$ is a time constant directly proportional to the filter circuit capacitance and inversely proportional to a user programmable bias current. Similarly, $\alpha$ is a gain term, tunable by additional user programmable bias currents (see [13] for a detailed explanation of the circuit details). The change in synaptic current $\Delta I_{syn}$ is therefore directly proportional to $I_w$ and to the pulse duration $\Delta T$ ($\Delta I_{syn} \propto I_w \cdot \Delta T$). However, while $I_W$ can be directly programmed by the on-chip current-mode bias generator (which produces the corresponding voltage bias $W$ to apply to $M_W$ of Fig. 1), the pulse duration $\Delta T$ can only be controlled indirectly, by programming different values of $I_L$ (i.e., the current flowing through $M_L$ of Fig. 1). In the particular implementation of the proposed circuit in the ROLLS chip, a capacitor $C_{pw}$ of $500\,fF$ was used.

In Fig. 3 we plot the voltage output of the log-domain pulse integrator in response to pulses extended by different values of $I_L$. As expected, smaller bias currents produce longer pulse durations. To show that longer pulse durations effectively increase the total synaptic current, we plot also the neuron's response to these pulses, for a constant $I_W$ bias, in Fig. 4. As the total synaptic current integrated by the silicon neuron circuit is larger for longer pulse durations, smaller settings of $I_L$ produce output spikes earlier: neuron response times to input pulses with large $\Delta T$ are shorter than response times to input pulses with short $\Delta T$.

In order to collect statistics and perform large sets of quantitative measurements on the variability of the neurons response times, we interfaced the ROLLS chip to an FPGA device and stimulated it using the AER communication protocol. We measured the time-stamps of the Address-Events both in input to the ROLLS (to stimulate the synapses) and in output (to measure the neuron response times). To evaluate the precise timing of the response-spikes relative timestamps of 40 ns resolution were attached to Address-Events produced by the chip. All communication and routing of events is managed in real-time by a soft processor implemented within the programmable logic fabric of the FPGA. The FPGA device then transmits the data to a host PC, via a USB link, for further processing and analysis.

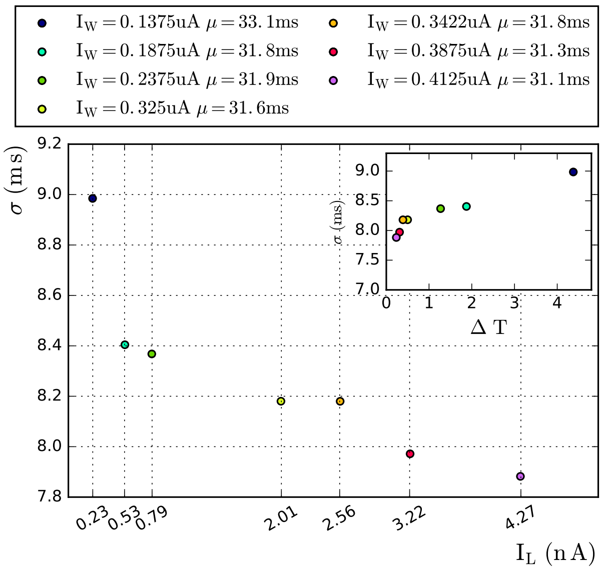

Fig. 5: Standard Deviation $\sigma$ of the neuron response times measured across the 256 neurons on the chip, for different settings of $I_L$ bias. Note that the neuron response times are kept approximately constant, by compensating for the changes in pulse duration with changes in synaptic weights. Inset is the same data, plotted over $\frac{1}{I_L}$ as it is proportional to $\Delta T$.

After the initialization of the ROLLS neuromorphic processor circuits with a proper set of analog biases, we configured the setup to stimulate a single synapse per neuron, in all 256 neurons in parallel. This can be achieved by broadcasting a single address-event across all rows of the synaptic array on the chip. We configured the synapses and neurons to produce an output spike in response to a single input address-event. At the onset of the experiment, the FPGA resets its time-stamp generator, and immediately after transmits the input address-event to the ROLLS chip. Neuron response times are measured over a time window of 1.5 s, and corresponding pairs of output neuron address and time-stamp relative to stimulation onset are measured. The experiment is repeated for different values of $I_L$ ranging from 0.25 nA to 5.99 nA. To compensate for the change in pulse duration, we also adjust the value of $I_W$ such that the neuron mean response time lies around 31 ms although the pulsewidth changes with $I_L$.

Figure 5 shows the standard deviation of the spiking output time of the neuron, in response to a single synaptic input, measured across the population of neurons on the chip, and repeated for different values of the leak current $I_L$, which is inversely proportional to the synaptic input pulse width. As expected, larger pulse widths (smaller $I_L$ settings) increase the variability across neuron responses, while smaller pulse

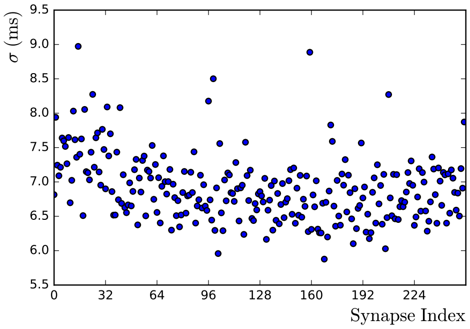

Fig. 6: Standard deviation of spike times $\Delta T$ produced by the population of neurons with $I_L = 2.01\,nA$, measured across the 256 synapses afferent to each neuron.

widths give rise to less variability, despite the fact that the neuron response times are approximately the same (thanks to the larger synaptic weights settings $I_W$ used to compensate for the shorter pulse widths).

Figure 6 shows the same measurement made for all 256 synapses, for the bias setting $I_W = 0.33\,\mu A$ and $I_L = 2.01\,nA$.

## IV. Conclusion

It has been shown that in recurrent networks of spiking neurons the ability to control stochasticity in the network dynamics allows to extend the network's learning features and memory storage-time [22]. Furthermore, manipulating the amount of stochasticity allows to adapt the system to the frequency at which patterns are presented [23]. The system we proposed exploits device mismatch to introduce stochasticity in the neurons firing response, by combining the use of minimum-size transistors for time dependent signals with the use of larger geometries for transistors that control the synaptic weights. We characterized how the amount of variability can be modulated by changing circuit configuration and biases, while maintaining the neuron's overall response properties constant. The developed set of tools used in this work, including the FPGA setup, provides a fast and reliable infrastructure for quantifying the variability present in the device under test through the measurement of the spike response times in response to single stimuli. We can now use this framework to investigate how the tunable variability of the neuron response times in the neuromorphic processor can be used to implement efficient random forest and bagging techniques for improving their classification accuracy, and to give rise to Stochastic behaviors in recurrent networks for implementing probabilistic neural models of computation and building networks able to carry out inference on their input data.

## Acknowledgment

This research has received funding from the European Union Seventh Framework Programme (FP7/2007- 2013) under grant agreement no. 612058 (RAMP), and from the e EU ERC program under grant grant agreement no. 257219 (NeuroP).

## References

[1] E. Chicca, F. Stefanini, C. Bartolozzi, and G. Indiveri, "Neuromorphic electronic circuits for building autonomous cognitive systems," *Proceedings of the IEEE*, vol. 102, no. 9, pp. 1367–1388, Sep 2014.

[2] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman, "How to grow a mind: Statistics, structure, and abstraction," *Science*, vol. 331, no. 6022, pp. 1279–1285, 2011.

[3] M. Mesulam, "Representation, inference, and transcendent encoding in neurocognitive networks of the human brain," *Annals of neurology*, vol. 64, no. 4, pp. 367–378, 2008.

[4] W. Ma, J. Beck, P. Latham, and A. Pouget, "Bayesian inference with probabilistic population codes," *Nature Neurosci*, vol. 9, no. 11, pp. 1432–1438, Nov 2006.

[5] E. O. Neftci, B. U. Pedroni, S. Joshi, M. Al-Shedivat, and G. Cauwenberghs, "Stochastic synapses enable efficient brain-inspired learning machines," *Frontiers in Neuroscience*, vol. 10, p. 241, 2016.

[6] W. Maass, "Noise as a resource for computation and learning in networks of spiking neurons," *Proceedings of the IEEE*, vol. 102, no. 5, pp. 860–880, May 2014.

[7] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[8] M. Skurichina and R. P. Duin, "Bagging, boosting and the random subspace method for linear classifiers," *Pattern Analysis & Applications*, vol. 5, no. 2, pp. 121–135, 2002.

[9] M.Rigotti, O. Barak, M. Warden, X. Wang, N. Daw, E. Miller, and S. Fusi, "The importance of mixed selectivity in complex cognitive tasks," *Nature*, May 2013.

[10] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S. K. Esser, R. Appuswamy, B. Taba, A. Amir, M. D. Flickner, W. P. Risk, R. Manohar, and D. S. Modha, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, Aug. 2014.

[11] A. Pavasović, A. Andreou, and C. Westgate, "Characterization of subthreshold MOS mismatch in transistors for VLSI systems," *Journal of VLSI Signal Processing*, vol. 8, no. 1, pp. 75–85, July 1994.

[12] T. Serrano-Gotarredona and B. Linares-Barranco, "Systematic width-and-length dependent CMOS transistor mismatch characterization and simulation," *Analog Integrated Circuits and Signal Processing*, vol. 21, no. 3, pp. 271–296, December 1999.

[13] N. Qiao, H. Mostafa, F. Corradi, M. Osswald, F. Stefanini, D. Sumislawska, and G. Indiveri, "A re-configurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128k synapses," *Frontiers in Neuroscience*, vol. 9, no. 141, 2015.

[14] P. Livi and G. Indiveri, "A current-mode conductance-based silicon neuron for address-event neuromorphic systems," in *International Symposium on Circuits and Systems, (ISCAS), 2009*. IEEE, May 2009, pp. 2898–2901.

[15] C. Rossant, D. Goodman, J. Platkiewicz, and R. Brette, "Automatic fitting of spiking neuron models to electrophysiological recordings," *Frontiers in Neuroinformatics*, pp. 1–14, 2010.

[16] R. Brette and W. Gerstner, "Adaptive exponential integrate-and-fire model as an effective description of neuronal activity," *Journal of Neurophysiology*, vol. 94, pp. 3637–3642, 2005.

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

[18] T. Delbruck, R. Berner, P. Lichtsteiner, and C. Dualibe, "32-bit configurable bias current generator with sub-off-current capability," in *International Symposium on Circuits and Systems, (ISCAS), 2010*, IEEE. Paris, France: IEEE, 2010, pp. 1647–1650.

[19] S. Deiss, R. Douglas, and A. Whatley, "A pulse-coded communications infrastructure for neuromorphic systems," in *Pulsed Neural Networks*, W. Maass and C. Bishop, Eds. MIT Press, 1998, ch. 6, pp. 157–78.

[20] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," in *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[21] D. Amit and S. Fusi, "Dynamic learning in neural networks with material synapses," *Neural Computation*, vol. 6, p. 957, 1994.

[22] N. Brunel, F. Carusi, and S. Fusi, "Slow stochastic hebbian learning of classes of stimuli in a recurrent neural network," *Network: Computation in Neural Systems*, vol. 9, no. 1, pp. 123–152, 1998.