Ultra-Low Power Silicon Neuron Circuit for Extreme-Edge Neuromorphic Intelligence

Arianna Rubino*, Melika Payvand*, and Giacomo Indiveri*
*Institute of Neuroinformatics, University of Zurich and ETH Zurich

Abstract—Recent years have seen an increasing interest in the development of artificial intelligence circuits and systems for cloud-less edge computing applications. In an effort to reduce power consumption even further, we propose beyond von-Neumann in-memory computing architectures that can process the signals at the sensor side using ultra-low power mixed-signal analog/digital circuits which have properly matched dynamics and time-constants. In this paper, we propose one of the main computing elements of such architectures, namely the silicon neuron, designed using analog circuits in an advanced FDSOI 22 nm node. Here we optimize the design of an Adaptive Exponential Integrate and Fire (AdExp IF) neuron model for producing neural dynamics with biologically plausible time constants. We explore the options of the 22 nm FDSOI technology to address the analog design issues that arise from advanced scaling (such as leakage) and minimize power consumption by using a novel current comparator circuit with current-driven positive feedback. We present circuit simulation results which reproduce biologically plausible responses and compare the circuit energy per spike with state-of-the-art architectures. The proposed neuron design consumes one order of magnitude less power compared to the state-of-the-art and two orders of magnitude less compared to a pure digital implementation.

Index Terms—Neuromorphic edge computing, Silicon neurons, FDSOI, Ultra-low power processing

I. INTRODUCTION

As the amount of data generated by the connected devices is ever more increasing, the power consumption for processing them is becoming more and more relevant. Local edge computing is gaining considerable attention because it promises to bring significant power savings by avoiding data transfer to remote (cloud) computing systems. In an effort to reduce power consumption even further, we propose to endow edge-computing sensory devices with ultra-low power processing circuits that can continuously monitor the data-streams directly on the sensing node, extract relevant information, and activate more powerful (and power hungry) computing or communication systems only when necessary. Recurrent spiking neural networks (SNNs) have been shown to be an ideal model for implementing such type of processing, provided their time constants are well matched to those of the signals of interest [1]. Mixed-signal event-driven neuromorphic circuits are natural candidates for implementing such SNN architectures and integrating them directly into IoT sensor nodes [2]. These SoC embedded sensory-processing systems can then run SNN artificial intelligence learning and inference algorithms in real-time on the sensed signals and dramatically reduce the bandwidth of their output signals. We denote this approach as “extreme-edge neuromorphic intelligence”. Such hardware is also suited for emerging memory technologies (e.g. memristors) and for the implementation of machine learning algorithms based on neural networks [3], [4]. However, mixed analog-digital design with deep sub-micron technology is challenging as a result of the increased leakage current that in advanced complementary metal-oxide-semiconductor processes becomes a significant portion of transistor’s ON current, leading to an increase in power consumption. Finally, as the technology node scales down and the transistor’s channel length decreases, its parameter variations (e.g. the threshold voltage) increase, and device mismatch increases even further.

In this paper, we present a sub-threshold neuron circuit that has been designed to implement large-scale multi-neuron multi-core neuromorphic computing architectures using a 22 nm Fully-Depleted Silicon on Insulator (FDSOI) process. We show how it is possible to implement bio-physically complex neural dynamics using ultra-low power compact analog circuits in advanced scaled processes, by analyzing the features of the 22 nm FDSOI technology and addressing the analog design issues that arise from the advanced scaling. To highlight the sub-parts of the 22 nm FDSOI silicon neuron circuits that are more sensitive to mismatch, we present Monte Carlo analysis results. Furthermore, we show how the use of an optimized current comparator with current-driven positive feedback significantly reduces power consumption.

II. MATERIALS AND METHODS

In 180 nm or even larger processes, transistors in sub-threshold regime usually operate with currents in the range of a few pico-Amperes to tens of nano-Amperes. In more advanced processes minimum-size transistors have considerably larger leakage currents. Therefore, to maintain the desired range of low currents, we performed circuit simulations of single transistors and determined their proper geometrical size. The devices available in the 22 nm FDSOI technology differ on the threshold voltage ($V_{th}$) value and hence on the $I_{off}$ as the two parameters are inversely proportional. Since our constraints are slow dynamics and low leakage, we considered the devices with high $V_{th}$, namely Ultra Low Leakage High Threshold Voltage Transistors (UHVT). As is illustrated in Fig. 1, by sweeping the width (80 nm - 600 nm) of both UHVT n-type and p-type with maximum length (36 nm), we determined the desired values of $I_{off}$. Based on these results, we designed an
Adaptive Exponential Integrate-and-Fire (AdExp IF) neuron circuit using UHVT MOSFETs provided by the 22 nm FDSOI technology. The circuit (Fig. 2) comprises an input Differential Pair Integrator (DPI) [5], [6] filter (P1 - N4), a current-based positive feedback module (P7 - N10), a current comparator (CC) block (P15 - N17), a spike reset circuit with refractory period functionality (P13 - N21), a spike generation inverter (P12 - N22) and a spike-frequency adaptation mechanism implemented with an additional DPI filter (see Fig.2(b)).

The input DPI models the neurons leak conductance, producing exponential sub-threshold dynamics in response to constant input currents. The integrating capacitance $C_{mem}$ (1 pF) represents the neurons membrane capacitance. The CC compares $I_{mem}$ with $I_{thr}$, set by an external bias, which sets the neuron’s spiking threshold. The positive feedback circuit models both sodium channel activation and inactivation dynamics, while the reset and refractory period circuit represents the potassium conductance functionality. The spike-frequency adaptation DPI models the neurons calcium conductance, and produces an after-hyperpolarization current ($I_{en, adp}$) proportional to the neurons mean firing rate. The neuron and the spike-frequency adaptation circuits are connected by a pulse extender which extends the spike duration.

To limit the Early effect, we used pseudo-cascode split-transistor sub-threshold technique, as done in [7]. This technique allows to generate bias currents on the order of pico-Amperes, necessary to have large time constants, while keeping the size of the capacitors to a minimum. For example, the diode-connected transistors N1-N2 in Fig. 2(a) and 2(b) are added to reduce the $V_{ds}$ of the transistors P3 and P4 respectively.

We used the same split-transistor sub-threshold technique for the bias current-mirrors, in order to compensate for mismatch, and to enhance current mirror operation to have precise control of small currents. Similarly, all the circuit parameters that require currents on the order of a few pico-Amperes have been implemented using transistors in series.

All capacitors are implemented using the Alternate Polarity Metal On Metal (APMOM) option. The value of the capacitance and the size are shown in Table I.

<table>
<thead>
<tr>
<th>Capacitance</th>
<th>Value</th>
<th>Width</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_{mem}$</td>
<td>1 pF</td>
<td>20 μm</td>
<td>20 μm</td>
</tr>
<tr>
<td>$C_{ahp}$</td>
<td>2 pF</td>
<td>28 μm</td>
<td>30 μm</td>
</tr>
<tr>
<td>$C_{ref}$</td>
<td>700 fF</td>
<td>16 μm</td>
<td>19 μm</td>
</tr>
<tr>
<td>$C_{pex}$</td>
<td>600 fF</td>
<td>15 μm</td>
<td>17 μm</td>
</tr>
</tbody>
</table>

III. RESULTS AND DISCUSSION

A. Circuit simulations

We optimized the design of the AdExp IF neuron for producing biologically plausible neural dynamics, with time constants matched to those of natural signals, such as speech or bio-signals.

Simulation results demonstrating examples of biologically plausible behaviors are shown in Fig. 3. Figure 3(a) shows the neuron spiking frequency versus input current (F-I curve), for different settings of the $I_{gain}$ bias. As expected, increases in $I_{gain}$ result in the increase of the neuron’s firing rate.
Figure 3(b) shows the neuron’s F-I curve for different $I_{ref}$ bias settings. As the $I_{ref}$ increases, the refractory period is shorter and hence the neuron’s maximum spiking frequency increases. Figure 4 demonstrates the spike-frequency adaptation behavior, obtained by appropriately tuning the relevant parameters in the AHP block of Fig. 2 and stimulating the neuron with a constant injection current.

The time constant of the spike-frequency adaptation circuit (63 ms) is twice the time constant of the neuron circuit (31 ms) since the capacitance $C_{ahp}$ (2 pF) is twice $C_{mem}$ (1 pF) using both $I_{leak}$ and $I_{tau\_adp}$ equal to 1 pA.

![Fig. 3](image)

![Fig. 4](image)

**B. Energy per spike**

Once proven that the design is able to reproduce a biologically plausible behaviour, we evaluated whether it can implement massively parallel large-scale neuromorphic processors. We compare the energy per spike of the neuron proposed in this work with previously proposed state-of-the-art neuromorphic processors in Table II.

<table>
<thead>
<tr>
<th>Work</th>
<th>[8]</th>
<th>[9]</th>
<th>[10]</th>
<th>[11]</th>
<th>This work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Techn.</td>
<td>180 nm</td>
<td>28 nm</td>
<td>180 nm</td>
<td>28 nm</td>
<td>22 nm</td>
</tr>
<tr>
<td>Type</td>
<td>Mixed</td>
<td>Mixed</td>
<td>Mixed</td>
<td>Digital</td>
<td>Mixed</td>
</tr>
<tr>
<td>$V_{dd}$</td>
<td>1.8 V</td>
<td>0.7 V-1 V</td>
<td>1.8 V</td>
<td>0.775 V</td>
<td>0.8 V</td>
</tr>
<tr>
<td>En./spike</td>
<td>883 pJ</td>
<td>2.3 nJ-30 nJ</td>
<td>10 pJ</td>
<td>800 pJ</td>
<td>990 fJ</td>
</tr>
</tbody>
</table>

The neuron designed in this work consumes one order of magnitude less compared to the most recent silicon neuron circuit design, the Sigma-Delta neuron proposed in [10].

**C. Monte Carlo Analysis**

We ran Monte Carlo simulations to evaluate the sensitivity of the circuit to mismatch. We performed this analysis with 500 runs for this neuron circuit, with DC current injected through P2 in Fig. 2, and with bias currents set to obtain a firing rate of approximately 65 Hz while switching off the spike-frequency adaptation circuit.

The Monte Carlo simulation produces a bi-modal distribution (see Fig. 5). The first mode gives a Gaussian distribution around the expected frequency value, 65 Hz. The second mode, shows that the neuron can fire at higher frequencies, around 180 Hz - 200 Hz. One possible reason for this could be that some parts of the circuit are faster than the refractory period part and make the neuron fire at higher frequencies exceeding the maximum value. In fact, the refractory period circuitry limits the neuron to fire with a maximum frequency of 100 Hz. If another part of the circuit is stronger than the refractory period, the node $V_{mem}$ does not reset completely staying at a higher voltage value. In this way, the neuron reaches the spiking threshold faster leading to higher spiking frequencies. We performed further Monte Carlo analysis to understand which specific transistors give this bi-modal distribution. Firstly, taking into account the mismatch sensitivity of the single neuron circuit. Secondly, considering also the sensitivity of the bias current mirrors. The result is reported in Fig. 6: In red the transistors that show sensitivity to mismatch by simulating the circuit without including the bias current mirrors and in orange the additional transistors that show sensitivity...
of the order of $10^{-13}$ - $10^{-12}$, which is considerably lower compared to an analogous neuron design implemented in a 180 nm CMOS process [8]. Furthermore analysis has shown how the proposed 22 nm implementation of the analog/digital IF AdExp neuron consumes two orders of magnitude less per spike than fully digital implementations. We studied with Monte Carlo analysis the mismatch sensitivity of the neuron circuit and highlighted the parts of the circuit that are most critical to be optimized for variations. In particular, we showed how the more sensitive sub-parts of the silicon neuron circuit are the input DPI filter, the first part of the CC, and the refractory period part. To minimize the effect of device mismatch we set the length and width of the most sensitive transistors to larger values.

**ACKNOWLEDGMENT**

This work is supported by the European Union’s Horizon 2020 ERC project NeuroAgent (Grant No. 724295), and Toshiba Corporation. This project has also received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826655. The JU receives support from the European Unions Horizon 2020 research and innovation programme and Belgium, France, Germany, Netherlands, Switzerland

**REFERENCES**