An FPGA-based Hardware-Efficient Fault-Tolerant Astrocyte-Neuron Network

Anju P. Johnson, David M. Halliday, Alan G. Millard, Andy M. Tyrrell and Jon Timmis
Department of Electronics, University of York, York YO10 5DD, UK
Email: {anju.johnson, david.halliday, alan.millard, andy.tyrrell, jon.timmis}@york.ac.uk

Junxiu Liu, Jim Harkin, Liam McDaid and Shvan Karim
School of Computing and Intelligent Systems, Ulster University, Derry BT48 7JL, UK
Email: {j.liu1, jg.harkin, lj.mcdaid, haji_karim-s}@ulster.ac.uk

Abstract—The human brain is structured with the capacity to repair itself. This plasticity of the brain has motivated researchers to develop systems which have similar capabilities of fault tolerance and self-repair. Recent research findings have proven that interactions between astrocytes and neurons can actuate brain-like self-repair in a bidirectionally coupled astrocyte-neuron system. This paper presents a hardware realization of the bio-inspired self-repair architecture on an FPGA. We also introduce a reduced architecture for an FPGA-based hardware-efficient fault-tolerant system. This is based on the principle of retrograde signaling in an astrocyte-neuron network by accelerating the calcium dynamics within the astrocyte. The hardware optimized implementation shows more than a 90% decrease in hardware utilization and proves an efficient implementation for a large-scale astrocyte-neuron network. An Average spike rate of 0.02 spikes per clock cycle were observed for both the proposed models of astrocytes in the case of 100% partial fault.

I. INTRODUCTION

Electronic systems are increasingly used in safety critical applications that demand low failure rates and fault tolerance. Fault tolerance has always been a standard feature of electronic systems targeted for long-life applications. Novel approaches to fault tolerance can be achieved by drawing inspiration from natural processes. Bio-inspired systems derive motivation from biological processes in the development of massively parallel computationally efficient systems. The human brain has remarkable computational capabilities, which can outperform standard computer systems in many real-world tasks. The brain is exceptional in its ability to self-repair, for example following stroke or injury. A number of researchers have demonstrated the potential of bio-inspired systems to achieve fault tolerance similar to a human brain [1]–[7]. It has recently been found that retrograde signaling via astrocytes has the capability to facilitate self-repair [8]. Astrocytes are star-shaped glial cells found in Central Nervous System (CNS) and Peripheral Nervous System (PNS) that enclose multiple synapses connected to a neuron. Astrocytes have the capacity to communicate in a feedback mode with the neurons, thereby establishing a bidirectional communication with the neurons forming a tripartite synapse that can self regulate neuronal activity [9].

In biological systems, independent units perform computation in parallel. For real world applications, this parallelism can be exploited to perform tasks orders of magnitude faster than

in software. FPGAs are preferred to GPUs for artificial neural networks applications [10], [11] due to their reconfigurability. Using FPGAs to implement artificial neural networks are an excellent choice because they combine computing capability, logic resources and memory capacity in a single device. In this work we employ FPGAs to implement the astrocyte-neuron based self-repairing unit. Designing any system to tolerate faults first requires the selection of a fault model—a set of possible failure scenarios along with an understanding of the depth and impact of each scenario. There can be various kinds of faults in a system when implemented in hardware including permanent faults (e.g. ageing faults, manufacturing defects, single event upsets) and temporary fault (e.g. power supply fluctuations, radiations). In this work we describe faults as a condition which results in a silent or near silent neuron caused by low transmission probability (PR) of a synapse, and repair is defined as the ability of the system to restore firing rates through an increase of PR. Our results demonstrate the ability of an FPGA-based implementation to exhibit self-repair under this fault condition.

The architectures proposed in this paper reduce the resource requirements for the astrocyte-neuron self-repair unit as follows:

1) Model-I of astrocyte: We have developed an improvised architecture for the self-repair unit by using approximation techniques for calcium dynamics of the astrocyte.
2) Model-II of astrocyte: We have developed a compact design methodology for implementing self-repair in the astrocyte-neuron network, by eliminating the complex processes occurring within the astrocyte. The developed architecture is 90% more resource efficient than [12].

The rest of the paper is organized as follows: Section II discusses the basics of the astrocyte-neural network model, followed by the proposed hardware implementation of the astrocyte-neuron self-repairing unit emphasizing a hardware-efficient astrocyte implementation in Section III. In Section IV, a reduced model of the self-repairing astrocyte-neuron network is presented. Section V presents our experimental results, and conclusions are drawn in Section VI.

II. ASTROCYTE-NEURAL NETWORK MODEL

The structure and working of the astrocyte-neuron self-repairing unit can be summarized as follows, in conjunction
The transmission probability \( (PR) \) of any synapse modulated by the effects of \( DSE \) and \( eSP \) is given by:

\[
PR(t) = PR(t_0)(1 + \frac{DSE + eSP}{100})
\]  

(1)

where \( PR(t_0) \) is the initial \( PR \) of the synapse. The sequence of events occurring at a tripartite synapse are briefly explained as follows: when a post-synaptic neuron fires, \( 2-AG \) is released. This is described by the following relation:

\[
\frac{d(AG)}{dt} = -\frac{AG}{\tau_{AG}} + \gamma_{AG}\delta(t - t_{sp})
\]  

(2)

where \( AG \) is the amount of \( 2-AG \) released. \( \tau_{AG} \) and \( \gamma_{AG} \) is the decay rate and the production rate of \( 2-AG \) respectively. \( t_{sp} \) is the time at which the post-synaptic neuron fires. The \( 2-AG \) released from the neuron give rise to a direct signaling at the tripartite synapse, leading to a reduction in \( PR \). The relation between the amount of \( 2-AG \) released and the \( DSE \) is modeled using a linear equation as follows:

\[
DSE = AG \times K_{AG}
\]  

(3)

where \( K_{AG} \) is a scaling factor used to balance equation (1).

In addition to direct signaling, The \( 2-AG \) released from the neurons associated with the astrocyte can also bind to the corresponding receptors (\( CB1R \)s) on the astrocyte. This leads to the production of \( IP_3 \) within the astrocyte.

\[
\frac{d(IP_3)}{dt} = \frac{IP_{3base} - IP_3}{\tau_{IP_3}} + \gamma_{IP_3}\sum_{i=1}^{q} AG
\]  

(4)

where \( IP_3 \) is the amount of \( IP_3 \) released within the astrocyte’s cytoplasm in response to \( 2-AG \). \( IP_{3base} \) is the baseline of \( IP_3 \) when the astrocyte is in steady state. \( \tau_{IP_3} \) and \( \gamma_{IP_3} \) is the decay rate and the production rate of \( IP_3 \) respectively. \( q \) is the number of neurons associated with an astrocyte.

In contrast to Wade et.al [8], we are using a parallel approach by considering the total \( 2-AG \) arriving at the astrocyte surface.

In response to alleviated \( IP_3 \), calcium builds up within the astrocyte. This is given by the following equation.

\[
\frac{d(Ca^{2+})}{dt} = J_{chan}(Ca^{2+}, h, IP_3) + J_{leak}(Ca^{2+}) - J_{pump}(Ca^{2+})
\]  

(5)

where \( J_{chan} \) is the calcium release dependent on \( IP_3 \) and \( Ca^{2+} \). \( J_{leak} \) models the leakage of \( Ca^{2+} \) out of the Endoplasmic Reticulum (ER) and \( J_{pump} \) is the amount of \( Ca^{2+} \) pumped into the ER. More details of astrocyte calcium dynamics are given in [13]. When the amount of calcium crosses certain threshold value, calcium spikes are formed, which in turn stimulate the release of glutamate (\( Glu \)). The amount of glutamate produced is given by:

\[
\frac{d(Glu)}{dt} = -\frac{Glu}{\tau_{Glu}} + \gamma_{Glu}\delta(t - t_{ca})
\]  

(6)

where \( Glu \) is the amount of glutamate discharged. \( \tau_{Glu} \) and \( \gamma_{Glu} \) is the decay rate and the production rate of glutamate respectively. \( t_{ca} \) is the time at which the calcium spikes. The glutamate released from the astrocyte binds to the corresponding receptors (group 1 \( mGluRs \)) on the pre-synaptic neuron.

Fig. 1: The tripartite synapse showing indirect and direct signaling of \( 2-AG \), figure from [8].
Fig. 2: Block diagram representing chain of biological processes involved in eSP and DSE production. (A) Model-I (B) Simplified Model-II of an astrocyte.

A probabilistic model is used for representing the synapse and is given by:

\[ I_{syn}^i(t) = \begin{cases} I_{inj}, & rand \leq PR, \\ 0, & rand > PR \end{cases} \]  (9)

A Linear Feedback Shift Register (LFSR) is used to generate a random number (rand). A fixed amount of current \( I_{inj} \) is injected into the synapse if the random number is less than or equal to the PR. The whole process involved is depicted in a block diagram shown in Fig. 2A.

III. FPGA-BASED ASTROCYTE NEURON NETWORK (MODEL-I)

The basic unit (Model-I) of a self-repairing astrocyte-neuron network is shown in Fig. 3. The architecture consists of two neurons (N1 and N2) and an astrocyte (A) shared between these two neurons. Each neuron is associated with a set of synapses (10 in our experiments). The synapses receive input spikes from other neurons. Any healthy synapse is assigned an initial PR of 0.5. For testing the basic unit, we used Poisson spike trains, which on an average, produce one spike per 4 clock cycles. In addition to the spike inputs, the synapses receive the direct signaling from the associated neuron and the indirect signaling from the associated astrocyte. The synapse processes these signals and makes a decision on the current to be injected onto the neuron. This is based on equations 1 and 9. In response to the total current injected, the neuron produces 2-AG leading to direct signaling. Also, an indirect signaling pathway is initiated via the astrocyte. The astrocyte processes the total 2-AG accumulated on its surface leading to indirect signaling through a set of complex processes represented in equation (4)–(7). Note that the eSP signal generated from astrocyte A is common to all synapses associated with neurons connected to it. The two signals (DSE and eSP) are balanced at the synapses so that the PR falls to 50% of the initial value once the system operates fully. The condition of a fault in a synapse is modeled by lowering the initial PR value to 0.

When a synapse fails, the current injected by it falls to zero as per equation 9. This lowers the 2-AG production and hence DSE falls, but the eSP signal exists in the system. Consider a case when all the synapses associated with N2 fail. Here, due to the presence of indirect pathway through neuron N1 and the astrocyte A, the PR of the synapses associated with N2 is raised leading to an increase in spiking activity of N2 after the fault.

The calcium dynamics for implementing the biological processes within the astrocyte requires complex multiplications and divisions as evident from [8] and [13]. Also, the biological parameters are precise in nature and require floating-point arithmetic for reproducing exact functionalities. A divider implemented in a Xilinx Intellectual Property (IP) core requires a latency of the order \( M + F \) (\( M \) is the width of dividend, and \( F \))
is the width of fractional reminder) clock cycles [15]. This implies that a chain of biological processes, at the astrocyte end, requires a higher latency (without pipelining). Additionally, implementing the floating point divisions involved in calcium dynamics of the astrocyte consumes an enormous amount of hardware resources. Hence when modeling an astrocyte process, the main operation to be approximated is division. In this work we approximate divisions involved in astrocyte calcium dynamics.

All the division operations involved in the calcium dynamics are approximated by sampling the constituent curves at regular intervals. The sampled values are stored in the FPGA. The sampled values are approximated by sampling the constituent curves at regular intervals. The sampled values are stored in the FPGA.

In this work we approximate divisions involved in astrocyte calcium dynamics.

All the division operations involved in the calcium dynamics are approximated by sampling the constituent curves at regular intervals. The sampled values are stored in the FPGA. The sampled values are approximated by sampling the constituent curves at regular intervals. The sampled values are stored in the FPGA.

IV. REDUCED ARCHITECTURE FOR BIO-INSPIRED SELF REPAIR (MODEL-II)

It has been found that astrocytes can encapsulate \( \sim 10^5 \) synapses and connect to multiple neighboring neurons [17], [18]. The biggest problem we face with the hardware realization of the classical model of astrocyte [8], incorporating the chain of complex chemical reactions, is the increased hardware footprint. We are interested in the accuracy and scalability of the neural networks. There are between 10 and 50 times more glia than neurons in the central nervous system of vertebrates [19] and particularly, astrocytes outnumber neurons by over fivefold [20]. To mimic brain-like functionality, the size of each astrocyte plays an important role. One of the main threats to the scalability of a network is the increased hardware utilization incurred in implementing the astrocyte. Considering scalability as an important metric, it is time to revisit the complex self-repairing astrocyte-neuron structure and reduce its complexity. Hence, the motivation behind the work presented in this paper is to simplify the existing architecture in terms of hardware, while retaining key features of direct negative feedback and the indirect positive feedback.

First, we simplify astrocyte equations (4)-(7) based on our observations, that imply the two features which underpin the spiking activity of a neuron are the balance between the direct and indirect signaling. In the reduced model, we simplify the astrocyte model by eliminating some activities within the astrocyte to reduce the complexity in building a self-repair unit. The modified block diagram is presented in Fig. 2B.

A in the self-repair architecture described in Section III, when a neuron is sufficiently depolarized, it produces 2-AG which leads to direct signaling. Equation (2) holds true for this model also. Equations (4)-(7) are merged to a single equation given by:

\[
\tau_{eSP} \frac{d(eSP_u)}{dt} = -eSP_u + m_{eSP} \sum_{i=1}^{q} AG_i \tag{10}
\]

where \( \tau_{eSP} \) is the decay rate of \( eSP \) and \( m_{eSP} \) is a weighting constant. \( q \) is the number of neurons associated with an astrocyte.

\[
eSP = K_2 \times eSP_u \tag{11}
\]

where \( K_2 \) is a scaling constant used to balance equation (1). By the action of \( DSE \) and \( eSP \) the \( PR \) depicted in equation (1) reduces to \( \sim 50\% \). Various parameters used in this reduced model (Model-II) are listed in Table I.

V. EXPERIMENTAL RESULTS

The hardware architectures shown in Fig. 2, for both astrocyte Model-I and Model-II, were designed using Verilog HDL. The designs were synthesized and implemented using Xilinx ISE 14.7 CAD software for the Xilinx Artix-7 target platform working at a clock frequency of 100MHz, and were simulated using Xilinx Isim. Power estimation of the circuits was carried out using Xilinx XPower Analyzer, and delay estimation using Xilinx Timing Analyzer. Xilinx Chipscope Pro was used for monitoring the activities of the system in the real hardware. We use the Euler method of integration with a fixed step size of \( \Delta t = 2^{-10} \). To represent floating point numbers we use a 32-bit representation where the first 16-bits represent the integer part, and the remaining 16-bits represent the decimal part. For example \( I_{inj} = 415.625 \times 10^{-6} \).

A. Implementation without a fault

In the first test case, all synapses had an initial \( PR \) (\( PR(t_0) \)) of 0.5, i.e. no fault condition (fault rate = 0%). The synapses associated with Neuron \( N1 \) and \( N2 \) receive both \( DSE \) and \( eSP \). Based on the input spike train and the \( PR \), a decision is made on the amount of current injected onto the neurons. When the input current is large enough to build up the membrane potential to a value higher than the threshold, the neuron outputs a spike. Fig. 4(A) shows the corresponding rise in the negative \( DSE \) signal for \( N2 \) (similar behavior is observed for \( DSE \) of \( N1 \)). Fig. 4(B) shows the \( eSP \) signal which is common to both \( N1 \) and \( N2 \). The \( PR \) of synapses associated with \( N2 \) is shown in Fig. 4(C). Fig. 4(D) shows the average spiking activity of the neuron \( N2 \). \( PR \) and spiking activity for \( N1 \) is similar to that for \( N2 \) (not shown in figure).

For determining the spiking activity, we use a window size of 128 clock cycles. In our model, the \( DSE \) and \( eSP \) signals achieve steady states within one millisecond. The hardware model we have developed therefore reaches a stable \( PR \) faster than the simulation model presented in [8], which achieves a steady state value after 100s. This is achieved by using an time scale faster than biology in our hardware implementation.

B. Implementation with fault

We implement a “faulty” synapse by lowering its initial \( PR \) to a value lower than 0.5. For instance, the initial \( PR \) of 80% of synapses associated with neuron \( N2 \) are intentionally lowered to 0 from 0.5 at 1000\( \mu \)s, i.e., 8 out of 10 synapses of neuron \( N2 \) are injected with a fault. Fig. 5(A)(Model-I) shows that \( PR \) of synapse-1, a faulty synapse, which initially undergoes a depression (at 1000\( \mu \)s), but tries to increase thereafter due to the repair. This behavior holds true for other faulty synapses also. The \( PR \) of healthy synapses of \( N2 \) increase to compensate for the faulty synapse as shown in Fig. 5(B). The \( PR \) of all synapses of neuron \( N1 \) remains unaffected as plotted in Fig. 5(C). Fig. 5(D) demonstrates the increase in spike frequency of \( N2 \) after the repair. Note: at 1000\( \mu \)s a dip in frequency is observed, and thereafter, due to the self-repairing capability, the spike rate increases. Fig. 5(E-H) shows same results for the reduced Model-II.
Fig. 4: Astrocyte-Neuron self repairing unit with no fault (A–D) for model-I (left) and (E–H) for model-II (right) of astrocyte. (A, E) DSE function of N2 (similar behavior is observed for DSE of N1). (B, F) eSP signal of both N1 and N2. (C, G) The probability of synapses connected to N1 and N2. (D, H) Firing rates of N2 (similar behavior is observed for firing rate of N1).

C. Implementation without an astrocyte

The system is tested in the presence and absence of an astrocyte. Fig. 6 plots the spiking activity of neuron N2 in the self-repairing unit neuron implemented on the FPGA. This is captured using Xilinx ChipScope CAD software. After some time, we induced a fault of 100% in neuron N2 (PR of all the synapses associated with neuron N2 are intentionally lowered to 0.25 from 0.5). The model containing an astrocyte induced with partial fault Fig. 6(A) shows spiking activity due to the repair initiated in the device. In the presence of an astrocyte with a catastrophic fault Fig. 6(B), the neuron N2 remains stagnant without producing any spikes. Also the PR associated with the faulty synapse is 0 in the presence of catastrophic fault, but has a moderate value (0.023 spikes per clock cycle) in the presence of partial fault. Similar behavior is observed for the reduced Model-II also (not shown). Faults of 70%, 80%, 90% and 100% were introduced to neuron N2 of the system (Fig. 7) (Model-I). It is evident that spiking activity is higher for the architecture with an astrocyte, thereby demonstrating the repair capability. It can be seen that the firing rate is higher in the presence of an astrocyte in all cases where faults are introduced. This behaviour holds true for both the models.
Fig. 5: *PR values of synapses and the firing rate of neuron N2 for a fault where 80% of synapses connected to N2 fail (PR lowered to zero) (A–D) for Model-I (left) and (E–H) for Model-II (right) of astrocyte. (A, E) *PR values of a faulty synapse of N2 (similar behavior is observed for all faulty synapses of N2), (B, F) *PR values of a non-faulty synapse of N2 (similar behavior is observed for all non-faulty synapses of N2), (C, G) *PR values of a synapse of N1 (similar behavior is observed for all synapses of N1), (D, H) Firing rates of N2 showing a spike rate falling after 1000 µs, and increasing thereafter due to the repair initiated by the astrocyte.

D. Power dissipation, delay and hardware resource overheads

Estimated power dissipation and delay of the overall architecture consisting of astrocytes of Model-I and Model-II are shown in Table II. Table III reports the hardware resource footprint of both models in comparison with the direct hardware implementation of [8]. As evident from these tables, the proposed architectures can be implemented with reduced hardware overhead, power dissipation and delay. Additionally, it is clear that Model-II is better for hardware-limited devices. Other hardware components (not shown in the paper) such as synapses, neurons and the input spike generators, do not contribute much to the hardware overhead.

VI. Conclusion

In this paper, a resource-efficient hardware architecture for an astrocyte-neuron self-repairing unit on an FPGA is presented. The present study helps to understand the main advantages of using hardware such as FPGAs for rapid prototyping.
Fig. 6: Chipscope results on real hardware (Nexys-4 Artix-7 FPGA board): Fault of 100% (PR of all the synapses associated with neuron N2 are intentionally lowered from 0.5) is induced on synapses connected to N2 at the trigger point T (A) Partial fault (PR(t_0) = 0.25) (B) Catastrophic fault (PR(t_0) = 0.0). The PR values of faulty synapse lowers to zero in the presence of catastrophic fault whereas has a moderate value in the presence of partial fault. Also the spikes produced by N2 ceases in the case of a catastrophic fault (B), whereas has a moderate value in the presence of partial fault (A).

TABLE I: Parameters used in self-repairing astrocyte-neuron network

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>V_\text{th}</td>
<td>9mV</td>
<td>1mV</td>
<td>1mV</td>
</tr>
<tr>
<td>R_m</td>
<td>1.261GΩ</td>
<td>1.2111Ω</td>
<td>1.2111Ω</td>
</tr>
<tr>
<td>\tau_{\text{mem}}</td>
<td>60ms</td>
<td>60ms</td>
<td>60ms</td>
</tr>
<tr>
<td>I_{\text{kin}}</td>
<td>6650pA</td>
<td>415.625mA</td>
<td>415.625mA</td>
</tr>
<tr>
<td>\gamma_{\text{AG}}</td>
<td>0.8µM</td>
<td>0.9999µM</td>
<td>0.9999µM</td>
</tr>
<tr>
<td>\tau_{\text{AG}}</td>
<td>10s</td>
<td>10s</td>
<td>10s</td>
</tr>
<tr>
<td>\gamma_{\text{FP}}</td>
<td>0.5µM</td>
<td>0.5µM</td>
<td>0.5µM</td>
</tr>
<tr>
<td>\tau_{\text{FP}}</td>
<td>7s</td>
<td>7s</td>
<td>7s</td>
</tr>
<tr>
<td>C_{\text{Gk}}</td>
<td>0.4</td>
<td>0.657</td>
<td>0.657</td>
</tr>
<tr>
<td>\gamma_{\text{Gk}}</td>
<td>10µM</td>
<td>0.825µM</td>
<td>0.825µM</td>
</tr>
<tr>
<td>\tau_{\text{Gk}}</td>
<td>100ms</td>
<td>1.61s</td>
<td>1.61s</td>
</tr>
<tr>
<td>\tau_{eSP}</td>
<td>40s</td>
<td>7s</td>
<td>7s</td>
</tr>
<tr>
<td>m_{eSP}</td>
<td>55 \times 10^3</td>
<td>3.5</td>
<td>0.21875</td>
</tr>
</tbody>
</table>

*Model-I uses same set of parameters for implementing calcium dynamics as in [8]

TABLE II: Power dissipation and delay overheads

<table>
<thead>
<tr>
<th></th>
<th>Model-I</th>
<th>Model-II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power (W)</td>
<td>0.339</td>
<td>0.173</td>
</tr>
<tr>
<td>Delay (ns)</td>
<td>37.343</td>
<td>15.792</td>
</tr>
</tbody>
</table>

TABLE III: Hardware utilization of astrocyte models

<table>
<thead>
<tr>
<th>Resource</th>
<th>Liu et.al [12]</th>
<th>Astrocyte Model-I</th>
<th>Astrocyte Model-II</th>
<th>% Decrease w.r.t [12]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slice</td>
<td>3552</td>
<td>3463</td>
<td>25</td>
<td>2.0956</td>
</tr>
<tr>
<td>Slice Reg</td>
<td>11666</td>
<td>231</td>
<td>32</td>
<td>98.0199</td>
</tr>
<tr>
<td>LUT</td>
<td>11316</td>
<td>11315</td>
<td>0</td>
<td>0.6740</td>
</tr>
<tr>
<td>BRAM</td>
<td>5</td>
<td>4</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>DSPs</td>
<td>42</td>
<td>42</td>
<td>4</td>
<td>90.4762</td>
</tr>
</tbody>
</table>

Fig. 7: Average firing rate of self repairing unit an astrocyte and without an astrocyte. It can be seen that the spike rate is higher in all cases in the presence of astrocyte.

and exploitation of inherent parallelism in brain-like self-repair mechanisms. Understanding the crucial features and extracting the required parameters are essential for mimicking the brain-like features for large scale designs. This implementation provides a stable PR at a faster instance over the prior art [8]. The results of FPGA implementations were in agreement with those of MATLAB simulations presented in [8].
VII. ACKNOWLEDGEMENTS

This work is funded by the EPSRC grant (EP/N007050/1).

REFERENCES


