A 31 pJ/Spike Hybrid Stochastic Neuromorphic Signal Processor

Amir Zjajo[1], *Nandish Mehta*,[2] *Rene van Leuken*[1]
[1]Circuits and Systems Group, Delft University of Technology, Delft, The Netherlands
[2]Berkeley Wireless Research Center, University of California, Berkeley, USA

Neuromorphic signal processing architectures capable of real-time applications are examined as a next generation, post-Moore, ultra-low-power computing solution [1]. Conventional von Neumann-type hardware (such as DSPs, GPUs, and FPGAs) in spiking neural networks require very high bandwidths (in the GHz range), and subsequently, high power dissipation, to efficiently transmit spike signal between the memory and the processor. In contrast, neuromorphic signal processing circuits are implemented on optimized, special purpose hardware, which can provide direct one-to-one mapping and low instruction execution redundancy [2]. Disparity between sequential-processing, conventional computing, and parallel, event-driven, biological neural systems is even more prominent in autonomous, real-time systems, especially in the presence of noisy and uncontrolled sensory input. In neural signal processing systems, the noise offers distinct advantages by inducing neuronal variability [3] and, successively, enhancing the sensitivity of neurons to environmental stimuli [4], inducing synchronization between neurons [5], and facilitating probabilistic inference [6]. Consequently, probabilistic noise models, as a resource for neural computation in the context of neuromorphic systems, are implemented as artificial neural networks and Boltzmann machines [7].

In this paper, proposed neuromorphic Boltzmann system performs robust neural computation using noise-induced stochastic equilibriums to regenerate static data distributions, and a neuromorphic core extended with self-learning and adaptation. The structure of the core consists of an input decoder that connects via 1024×256 programmable synapses to 256 integrate-and-fire neurons, an I/O network communication layer, and an activity-dependent dynamic voltage and frequency scaling (DVFS) circuits for active power reduction. The neuron circuits are current-mode, conductance-based, compact, process input data on demand, in real time, and produce fast asynchronous digital output pulses. The neuron's time constants and spike frequency adaptation are controlled with adaptable circuit biasing, and consequently, the circuit can generate a wide range of time constants and spiking behavior. The neuron circuit employs positive feedback to reduce the neurons' switching time, and reduce consumed power. The noise-induced stochastic dynamics [8] are implemented with log-domain subthreshold circuits, which offer high energy-efficiency and minimal power-delay products over several decades of operating range.

A wide range of neural network algorithms, transformed into a hardware compatible format, can be implemented in the proposed neuromorphic core. We examined the utility of the noise-induced stochastic dynamics of Boltzmann machines in generalizing the variability of EEGs with a test dataset containing the segments of 10 normal and 10 abnormal EEG recordings from the human neocortex and basal ganglia. The training data were extracted from a 10-minute long recording in MIT-BIH database. Experimental results show that the proposed system can distinguish EEG signals with 96 % accuracy, within the early onset of 0.1 ms, due to the recursive accumulation of the signal difference, and concurrent implementation of the Boltzmann architecture arrays. The neuromorphic core is fully re-configurable, and consumes only [9]-[10] 31 pJ/spike at 0.8 V supply voltage in 65 nm CMOS technology.

References

[1] E. Chicca, *et al.* "Neuromorphic electronic circuits for building autonomous cognitive systems," *Procedings of IEEE*, vol. 102, no. 9, pp. 1367-1388, 2014.
[2] C. Zamarreno-Ramos, *et al.* "On spike-timing-dependent-plasticity, memristive devices, and building a self-learning visual cortex," *Frontiers in Neuroscience*, vol. 5, pp. 1-36, 2011.
[3] G.B. Ermentrout, R.F. Galan, N.N. Urban, "Reliability, synchrony and noise," *Trends in Neuroscience*, vol. 31, no. 8, pp. 428-434, 2008.
[4] K. Wiesenfeld, F. Moss, "Stochastic resonance and the benefits of noise: From ice ages to crayfish and squids," *Nature*, vol. 373, pp. 33-36, 1995.
[5] J. M. Casado, "Synchronization of two Hodgkin-Huxley neurons due to internal noise," *Physics Letters A*, vol. 310, nos. 5-6, pp. 400-406, 2003.
[6] W.J. Ma, *et al.* "Bayesian inference with probabilistic population codes," *Nature Neuroscience*, vol. 9, no. 11, pp. 1432-1438, 2006.
[7] B.U. Pedroni, *et al.* "Neuoromorphic adaptations of restricted Boltzmann machines and deep belief networks," *Proceedings of IEEE International Joint Conference on Neural Networks*, pp. 1-6, 2013.
[8] A. Zjajo, C. Galuzzi, R. van Leuken, "Stochastic noise analysis of neural interface front end," *Proceedings of IEEE International Symposium on Circuits and Systems*, pp. 169-172, 2015.
[9] P. Merolla, *et al.* "A digital neurosynaptic core using embedded crossbar memory with 45pJ per spike in 45nm," *Proceedings of IEEE Custom Integrated Circuits Conference*, pp. 1-4, 2011.
[10] V. Narayanan, *et al.* "Video analytics using beyond CMOS devices" *Proceedings of IEEE Design, Automation and Test in Europe Conference*, pp. 1-5, 2014.

Research reported in this publication was supported in part by the European Union and the Dutch government, as part of the CATRENE program under Heterogeneous INCEPTION project.

A 31 pJ/SPIKE HYBRID STOCHASTIC NEUROMORPHIC SIGNAL PROCESSOR

Amir Zjajo[1], Nandish Mehta[2], Rene van Leuken[1]

[1]Delft University of Technology, The Netherlands, [2]University of California, Berkeley, USA

TU Delft
Delft University of Technology

Abstract

- Neuromorphic signal processing architectures are examined as a next generation, post-Moore, ultralow-power computing solution.
- Conventional, sequential-processing, von Neumann-type hardware (e.g., DSPs, GPUs, FPGAs) in spiking neural networks require very high bandwidths (in GHz range), and subsequently, high power dissipation, to efficiently transmit spike signal between the memory and the processor.
- Parallel, event-driven, neuromorphic signal processing circuits are implemented on optimized, special purpose hardware.
- In this project, we designed neuromorphic core for robust neural calculation that includes: the I/O network communication layer, hybrid analog/digital processing units, and activity dependent DVFS.
- Each processing unit incorporates integrate-and-fire neurons and synapses, wherein memory and computation are co-localized.
- The neuromorphic core is fully reconfigurable, and consumes only 31 pJ/spike at 0.8 V supply voltage in 65 nm CMOS technology.

Background

- Silicon neurons consist of one or more synapse blocks, and a soma block.
- The synapse blocks receive spikes from other neurons, integrate them over time, and convert them into currents.
- The soma block performs the spatio-temporal integration of the input signals, and generates the output analog action potentials and/or digital spike events.
- The soma block can be further subdivided into several functional blocks that reflect the computational properties of the theoretical models they implement (i.e., temporal integration block, a spike generation block, a refractory period block, and a spike-frequency or spiking threshold adaptation block).

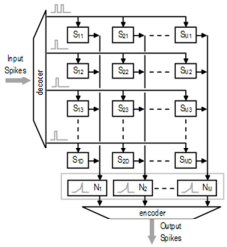
Neuromorphic Electronic Circuits

- Emulating the style of computation of the brain using the physics of silicon to reproduce the biophysics of the neural tissue:
  - VLSI devices for building real-time sensory-motor systems
  - Compact and low-power behaving systems
  - Brain-machine interfaces
  - Neural computation.
- Exploring the computational properties of the neural system they emulate, and gain a better understanding of its operational principles:
  - Spike-timing in the computational neuroscience
  - Event-driven computing systems.

Neuromorphic Signal Processor

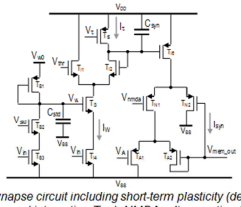
- Neuromorphic Boltzmann system
  - Complete robust computation using noise-induced stochastic equilibriums to regenerate static data distributions
  - Neuromorphic core extended with self-learning, self-organization and adaptation.
- Neuromorphic core consists of parallel arrays of processing units, and incorporates integrate-and-fire neurons and synapses.
  - The core comprises of input decoder that connects via 1024×256 programmable synapses to 256 neurons
  - The neuron circuits are current-mode, conductance-based, compact, and process input data on demand, in real time.

Overview of the Neuromorphic Core



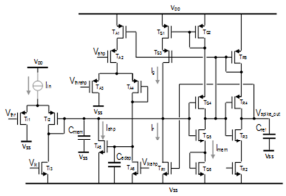
<Internal blocks of the neuromorphic core with dedicated synapse. The topology relaxes the dynamic range conditions on the inputs.>

- Synapses – Fundamental elements for computation and information transfer in neural systems
  - Short-term plasticity mechanisms are effective tools for processing temporal signals.



<Synapse circuit including short-term plasticity (depression T_{S1-3}, and integration T_{I1-6}), NMDA voltage gating (T_{N1-2}), and conductance-based functional blocks (T_{A1-2}). [1]>

An Adaptive Neuron Circuit

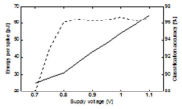


<Adaptive neuron circuit including: i) input integrator block T_{I1-3}, ii) spike-frequency adaptation circuit T_{A1-6}, iii) spike event generation circuit T_{G1-6}, and iv) digital communication and reset block T_{R1-6}. [2-3]>

- Computational simplicity and compactness of generalized integrate-and-fire neuron models make them valuable options for VLSI implementations.
- Adaptive neuron circuit is compact, low power, implements refractory period and spike-frequency adaptation, and has biologically realistic time constants.

Hardware Evaluation

- The neuron circuit can produce a wide range of time constants and spiking behavior, and can produce fast asynchronous digital output pulses.
- Noise-induced stochastic dynamics is implemented with log-domain subthrashold circuits, which offer high energy-efficiency and minimal power-delay products over several decades of operating range.
- A test dataset containing the segments of 10 normal and 10 abnormal EEG recordings from the human neocortex and basal ganglia. The training data were extracted from a 10-minute long recording in the MIT-BIH database.
- The system can distinguish EEG signals with 96% accuracy, within the early onset of 0.1 ms.
- The power-efficient solution is achieved by a combination of the algorithm and circuit techniques, i.e. supervised fine-tuning with global weighting, concurrent hybrid analog/digital implementation of the processing units, dynamic voltage and frequency scaling techniques.
- The circuit Implemented in 65 nm CMOS technology, consumes 31 pJ/spike at 0.8 V supply voltage (for 96 % accuracy).



<Energy per spike versus supply voltage, and classification accuracy.>

Neuromorphic Boltzmann System

- Noise-induced stochastic dynamics of restricted Boltzmann machines (RBM) examined in generalizing the variability of EEGs
- Unsupervised training of a single binary RBM via stochastic gradient descent:

$$d(P^0 \,\|\, P^\infty;\Theta^l) := \sum_{\mathbf{v}^l} P^0(\mathbf{v}^l)\log\left(\frac{P^0(\mathbf{v}^l)}{P^\infty(\mathbf{v}^l;\Theta^l)}\right)$$

assuming system states $\mathbf{s}^l:=[(\mathbf{v}^l)^T,(\mathbf{h}^l)^T]^T$, with distribution

$$d(P^0 \,\|\, P^\infty(\mathbf{v}^l;\Theta^l) = \frac{1}{Z(\Theta^l)}\sum_{\mathbf{h}^l}\exp(-H(\mathbf{s}^l;\Theta^l))$$

given network parameters $\Theta^l:=(\mathbf{W}^l,\mathbf{b}^l)$, i.e. the weights $\mathbf{W}^l$ and biases $\mathbf{b}^l$. The function H reads as

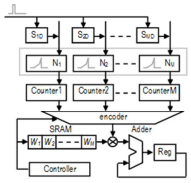
$$H(\mathbf{s}^l;\Theta^l) := -(\mathbf{v}^l)^T\mathbf{W}^l\mathbf{h}^l - (\mathbf{b}^l)^T\mathbf{s}^l$$

Gaussian-distributed states for the input RBM are introduced through a quadratic energy function. Linear output layer on input sample x is defined as

$$\mathbf{v}^{out}(\mathbf{x}) = \mathbf{W}^{OUT}\mathbf{h}(\mathbf{x}) + \mathbf{b}^{OUT}$$

Notes: Output RBM learns input-output associations leading to an indirectly maximized discriminant criterion.

- Hardware implementation



<A section of the classifier architecture including SRAM, adder, multiplier, register and controller.>

Summary

- The neuromorphic signal processor designed for robust neural calculation, consisting of the I/O network communication layer, programmable hybrid analog/digital processing units, and activity dependent DVFS.
- Noise-induced stochastic dynamics of RBM examined in generalizing the variability of EEGs.
- The circuit is Implemented in 65 nm CMOS technology and consumes only 31 pJ/spike at 0.8 V supply voltage.

References

[1] C. Bartolozzi, et al., Neur. Comp., 2007.

[2] R. Jolivet, et al., J. Neurophys., 2004.

[3] R. Naud, et al. , Front. Neur. Conf., 2009.