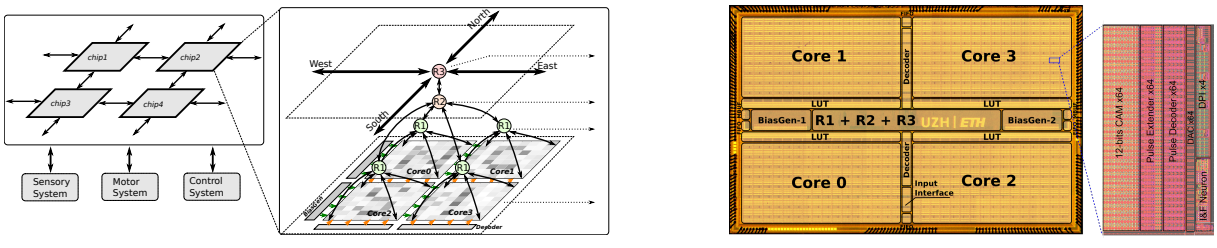
Robust implementation of cognitive computing computational primitives in mixed-signal neuromorphic processors

Dmitrii Zendrikov, Sergio Solinas, Ning Qiao and Giacomo Indiveri
Institute of Neuroinformatics, University of Zurich and ETH Zurich, Switzerland

Summary. Hardware implementations of brain-inspired cognitive computing architectures are typically characterized with constraints and limitations that are significantly different from those found in conventional processor designs. In particular, mixed-signal analog-digital neuromorphic processors comprise populations of silicon neurons and dynamic synapses that can faithfully emulate the physics of computation of cortical circuits. As a consequence, the computing elements in these devices are often affected by the same limitations observed in biological neurons and synapses, i.e., low resolution, slow speed, high variability and high sensitivity to noise. In this work, we propose spiking neural network configurations that implement robust computational primitives for building complex real-time neural information processing systems. We validate these architectures on the Dynamic Neuromorphic Asynchronous Processor (DYNAP) device and demonstrate their cognitive computing abilities with decision making winner-take-all circuits coupled in a relational network.

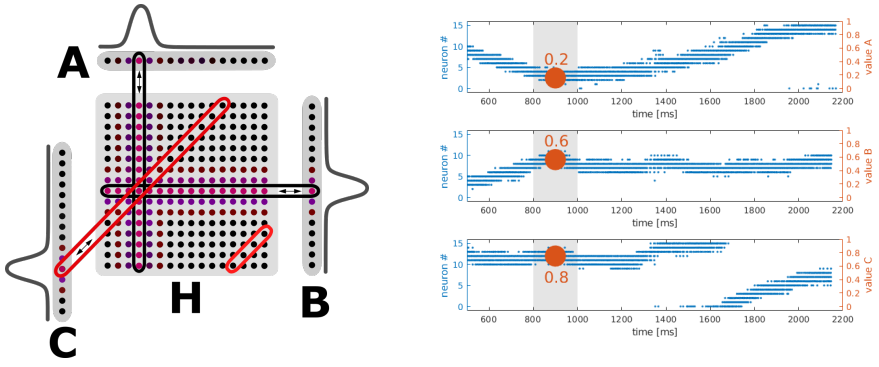
Novel brain-inspired processing algorithms and artificial neural networks are showing remarkable results in a wide variety of data processing architectures [4]. However, implementing such algorithms on conventional von Neumann computing systems is not ideal, as their time-multiplexing mode of operation and memory bottleneck [1] impose very stringent requirements on their power budget. Recently a new class of dedicated hardware devices, optimized for implementing neural network algorithms, started to appear. Among these, the architectures that offer the highest power-savings for a given number of operations are the ones designed to implement *spiking* neural networks (SNNs). While there are by now well established design workflow and programming frameworks for specifying the structure and parameters of rate-based artificial neural networks to solve a wide range of tasks, designing and configuring SNNs to carry out desired functions is still an open challenge. Solving this challenge is particularly important in dynamic autonomous agent scenarios that require real-time processing of sensory signals and production of desired motor sequence commands. If these autonomous agents are realized as small and compact robotic platforms or small intelligent sensors that need to transmit data to further processing stages only when necessary (e.g., indoor autonomous drones or environmental sensors), then it is crucial that these computational primitives are compatible and well-matched to ultra-low power and compact electronic substrates that can implement them.

Here we present an example of a cognitive task that makes use of basic SNN computational primitives which can be implemented using low-precision and noisy silicon neurons. The goal is to create a neural network that can realize basic relations dynamically linking three variables (A, B, and C). The network proposed is capable of performing cue-integration and inferring the value of any of the three variables given the other two, a task known as omni-directional function approximaton [2]. The basic computational primitive that is used to perform this task is a spiking neural Winner-Take-All (WTA) network. Such WTA networks have been shown to be able to implement computational primitives with both analog and digital processing characteristics [3]. Furthermore, it has been demonstrated that such networks can be successfully used to synthesize robust cognitive computing state-machines on mixed-signal neuromorphic devices [6]. Following up on these achievements, we demonstrate the implementation of the proposed relational network using the Dynamic Neuromorphic Asynchronous Processor (DYNAP): a recently developed mixed signal analog-digital neuromorphic processor which is among the most efficient



(a) **(b)**

Figure 1 DYNAP architecture. (a) Multi-chip and multi-core arrangement, with hierarchical routing scheme. (b) Die photograph with highlight of single neuron block.



(a) (b)

Figure 2 Example of a three-way relation network encoding A + B = C. The network allows omni-directional inference of a missing input provided the other two. Variable values are encoded by the positions of population activation peaks in WTA networks. (a) The relation is encoded by the highlighted connectivity patterns. Each neuron of three 1D WTA populations, A, B, and C (16 excitatory neurons each, inhibitory neurons not shown), is bidirectionally connected to a corresponding row, column or diagonal of neurons in a 2D WTA hidden population H (256 neurons). Connections between populations C and H include a y-axis wrap-around to maintain symmetry of a total number of connections. (b) Raster plot showing the spiking activity of populations A, B, and C. The experiment shows inference of C given A and B, as they change over time. The shaded area shows a moment in time corresponding to example values used in (a). Note the wrap-around of C after 1600 ms.

devices available in terms of number of synaptic operations per second per Watt [5]. A block diagram and chip micro-graph of the DYNAP is shown in Fig. 1. It comprises four cores of 256 neurons each, implemented as adaptive exponential integrate and fire neurons, with 64 dynamic synapses per neuron. The weight resolution of a single synapse is of 2 bits, but source neurons can target multiple synapses of the destination neuron to increase their effective resolution. A detailed description of the DYNAP is provided in [5].

In the relational network proposed the variable values are encoded by three distinct WTA networks, while the specific relation it implements is hard-coded in the connectivity of a hidden population (H; Fig. 2a). In spite of the noisy nature of the silicon neurons, the hardware setup is able to provide a stable inference. Additionally, time-varying inputs produce proper inference of the free variables (see Fig. 2b). As this network can infer its output by integrating its inputs, it can be used as higher-level computational primitive that can be connected to other networks, to provide a modular hardware building block for composing more complex networks, i.e. a functional node of a larger neural structure. From the cognitive point of view, the omni-directional relation features of this network constitute the neural correlate of the basic decision making processes which are used in neural systems to compose more complex cognitive functions.

References

[1] J. Backus. Can programming be liberated from the von Neumann style? A functional style and its algebra of programs. *Communications of the ACM*, 21(8):613–641, 1978.

[2] S. Deneve, P. Latham, and A. Pouget. Efficient computation and cue integration with noisy population codes. *Nature Neuroscience*, 4(8):826–831, 2001.

[3] R. Douglas and K. Martin. Recurrent neuronal circuits in the neocortex. *Current Biology*, 17(13):R496–R500, 2007.

[4] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[5] S. Moradi et al. A scalable multicore architecture with heterogeneous memory structures for dynamic neuromorphic asynchronous processors (DYNAPs). *Biomedical Circuits and Systems, IEEE Transactions on*, pages 1–17, 2017.

[6] E. Neftci et al. Synthesizing cognition in neuromorphic electronic systems. *Proceedings of the National Academy of Sciences*, 110(37):E3468–E3476, 2013.