Real-time inference in a VLSI spiking neural network

Dane Corneil*, Daniel Sonnleithner*, Emre Neftci*, Elisabetta Chicca†, Matthew Cook*,
Giacomo Indiveri*, Rodney Douglas*
*Institute of Neuroinformatics
University of Zurich and ETH Zurich
Email: emre@ini.phys.ethz.ch
† Cognitive Interaction Technology - Center of Excellence
Bielefeld University, Germany

***Abstract*—The ongoing motor output of the brain depends on its remarkable ability to rapidly transform and fuse a variety of sensory streams in real-time. The brain processes these data using networks of neurons that communicate by asynchronous spikes, a technology that is dramatically different from conventional electronic systems. We report here a step towards constructing electronic systems with analogous performance to the brain. Our VLSI spiking neural network combines in real-time three distinct sources of input data; each is place-encoded on an individual neuronal population that expresses soft Winner-Take-All dynamics. These arrays are combined according to a user-specified function that is embedded in the reciprocal connections between the soft Winner-Take-All populations and an intermediate shared population. The overall network is able to perform function approximation (missing data can be inferred from the available streams) and cue integration (when all input streams are present they enhance one another synergistically). The network performs these tasks with about 80% and 90% reliability, respectively. Our results suggest that with further technical improvement, it may be possible to implement more complex probabilistic models such as Bayesian networks in neuromorphic electronic systems.**

I. Introduction

The combination of sensory input cues and the inference of missing information from noisy, incomplete sensory cues are fundamental computations carried out by the brain [1]. The brain performs these prodigious feats using networks of slow, spike-communicating neurons, which stand in stark contrast to the technology and algorithms employed by digital processors. Recent theoretical studies have demonstrated with simulations how such computations could be performed in a biologically plausible manner, using arrays of ideal neurons with real-valued outputs [2]–[4].

Our contribution here is to demonstrate that this biological style of sensory processing can be emulated physically and in real-time using networks of neuromorphic Very Large Scale Integration (VLSI) [5] neurons that communicate by asynchronous events (spikes). Our neuromorphic network is able to combine and transform sensory input cues into outputs according to an arbitrary user-specified function. It consists of four networks of spiking neurons, three of which provide sensory input while the fourth combines these inputs. The three input networks are configured as 1-D arrays of Integrate-and-Fire (I&F) neurons whose lateral excitatory and global inhibitory couplings implement a soft Winner-Take-All (sWTA) network [6]. These input sWTA populations provide place-encoded representations of their sensory variables. The relationship between the variables is specified by the bi-directional connectivity between the individual sWTA networks and a shared intermediate 2-D WTA network.

The recurrent excitation developed by these networks, constrained by the patterns embedded in their interconnections, provides the gain necessary to recover an unspecified cue when the two others are specified (function approximation), or to sharpen the response profile of each variable by integrating the information between the specified cues (cue integration) [7]. The four populations form an attractor network, where the state of each is constrained by the function relating them. The recurrent pattern of connections in our network is consistent with the observed connectivity of the neocortex [8] and has been proposed as an important neural computational primitive [9] [10] that can be combined easily and stably in large networks [11].

Our demonstration is of interest for three reasons. The network is a step towards configuring neuromorphic asynchronous spiking systems to perform generic processing. As well, the interaction of populations of neurons with various profiles of activation resembles the interaction of statistical distributions used in graphical probabilistic models such as Bayesian networks. Finally, the realization in distributed electronic hardware may offer a scalable technology for real-time data integration and inference.

II. The Multi-Chip Setup

The network is distributed across two different types of chips: a 2-D Integrate-and-Fire (IF2D) chip, which implements a 2-D array of I&F neurons, and a 2-D Winner-Take-All (WTA2D) chip, which implements a current-mode Winner-Take-All (WTA) network. These are neuromorphic chips which receive and transmit spikes using the Address Event Representation (AER) communication protocol [12]. The asynchronous infrastructure used to transmit spikes across chip boundaries makes use of dedicated AER communication and mapper FPGA boards, which allow the user to specify arbitrary network connectivity schemes [13].

A. The 2-D Integrate-and-Fire (IF2D) Chip

The IF2D consists of a 2-D sheet of 32×64 neurons, with three externally addressable AER synapse circuits each (2

excitatory, 1 inhibitory). The synapses are pulse integrators that can be stimulated by other neurons on the same chip, or by Address-Events from outside sources to produce biophysically realistic Excitatory Post-Synaptic Currents (EPSCs). In addition, there are local hard-wired synapse circuits that integrate spikes from nearest neighbor neurons on the same chip. There is no dedicated locally hardwired inhibitory neuron pool. Instead, the global inhibition required for sWTA operation can be provided by making an all-to-all mapping from the neurons in the array to their AER inhibitory synapses, via the AER mapping infrastructure. By activating the local recurrent excitatory connections and external recurrent inhibitory connections to provide global inhibition, it is possible to produce the sWTA function, in which the winners are a small population of neighboring neurons that suppress or reduce the activity of all other neurons in the network. The chip can be configured to provide up to 32 independent sWTA networks consisting each of (up to) 64 neurons. The IF2D was fabricated using a standard AMS 0.35 μm CMOS process, and covers an area of about 15 mm^2.

B. The 2-D Winner-Take-All (WTA2D) Chip

The WTA2D chip comprises a grid of 32×32 cells. Each cell contains an AER excitatory input synapse, a current-mode WTA circuit, and an output I&F neuron [14]. The AER excitatory synapses integrate the input spikes and produce an output current which is proportional to the frequency of the input spike train. The current-mode WTA circuit then selects the cell receiving the strongest input and activates its corresponding output neuron. Output Address-Events therefore encode the position of the winning cell. Due to the nature of the current-mode WTA circuit, only one neuron is active at any given time. The WTA2D chip was fabricated using a standard AMS 0.35 μm CMOS process and covers an area of about 10 mm^2.

C. Network Architecture

The organization of the network is illustrated in Fig. 1. We selected three distinct arrays of 32 neurons from the IF2D chip. These arrays formed the populations whose activities represent three cues: X, Y and Z. We configured each array as a sWTA network, by activating the local recurrent first- and second-neighbor excitatory connections and the global inhibitory ones.

We then established bidirectional excitatory connections between each of the three populations X, Y and Z and the neurons in the WTA2D chip (population R). Specifically, the R neurons in the WTA2D chip were connected to the X, Y and Z neurons in the IF2D chip via their first AER excitatory synapse. Each neuron in X was connected to every neuron in the corresponding *column* on R; the neurons in Y were connected to the *rows* of R. The connection matrix from R to Z determines the user-specified function to implement. In the case of Fig. 1 the relational function is the *mean* ($Z = \text{floor}[(X+Y)/2]$). For instance, cell $[2,4]$ on R was connected to neuron 3 on Z. All connections between R and X, Y, and Z, also involved the neuron's nearest neighbors. For instance, neuron 4 on X was also connected to columns 3 and 5 on R.

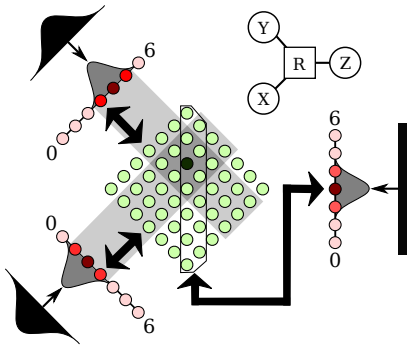

Fig. 1: Illustration of the network architecture applied to a function approximation experiment. In this example the user-defined function is $Z = \text{floor}[(X+Y)/2]$. Each neuron in populations X and Y feeds into all neurons in the associated diagonal row on population R. Neural firing activity is indicated by color intensity. According to the user-defined relation, each neuron in Z is connected to two adjacent vertical columns on R. External input signals are represented by the left and right black shapes. On the left, Gaussian inputs centered around neuron 2 and 4 are applied to X and Y respectively. A constant input is applied to Z. The firing activity in X and Y produces maximum activity at location $[2,4]$ on R; this cell is in a vertical column connected to neuron 3 on Z.

We generated input signals on a PC and provided them to the X, Y and Z populations on the IF2D chip by stimulating the target neurons via their second AER excitatory synapse. To reduce the effect of fabrication mismatch, we calibrated the inputs to neurons in populations X, Y and Z using the method described by Neftci et al. [15]. The neurons in the WTA2D chip did not receive any external stimulation from the PC.

III. Real-time experiments

There are at least two types of computations that such network architectures can achieve [2]. The first is the recovery of an unspecified cue when the other two cues are specified (function approximation); the second is the sharpening of the response profiles of the variables by integrating the information between the specified cues (cue integration). The two experiments performed demonstrate both of these computations.

For both experiments, we carried out 30 trials using the arbitrarily chosen function $Z = \text{floor}[(X+Y)/2]$. To evaluate the sWTA network effects we also carried out trials with identical inputs, but with only feed-forward connections active and all local WTA connections inactive (*feed-forward* condition), as well as trials in which only the connections to the intermediate WTA2D layer were inactive (*sWTA only* condition).

A. Function Approximation

We applied regular spike trains with Gaussian spatial distributions (75 Hz maximum firing rate, $\sigma = 1.75$) to neurons of populations X and Y. Inputs were centered on random positions (at least 3 neurons from the edges to avoid boundary effects caused by the lack of recurrent connections wrapping around the neural arrays). Poisson-distributed spike trains

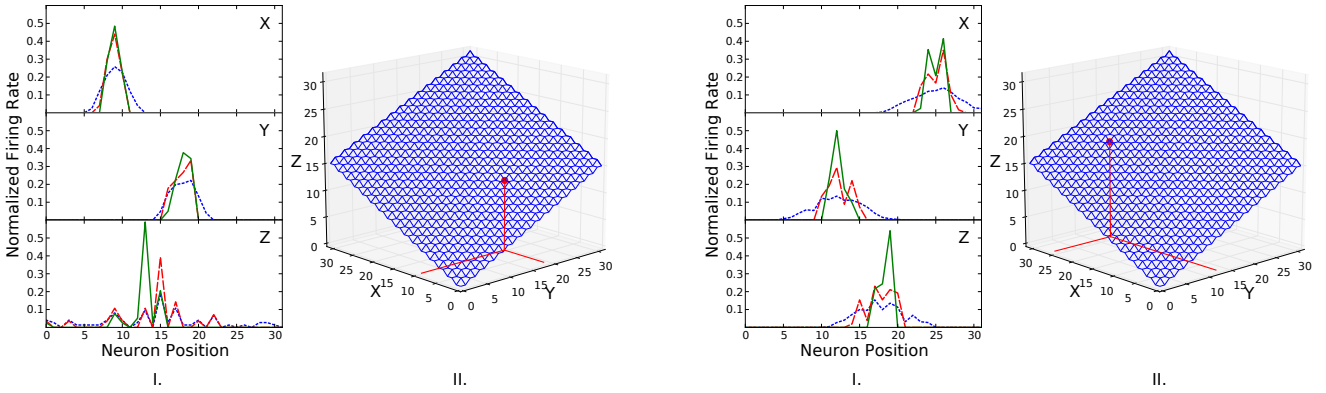

Fig. 2: Function Approximation (a) and Cue Integration (b) experimental results. I. Normalized firing rates of populations X, Y and Z measured from 700 ms to 1000 ms (green solid traces for real output, red broken traces for *sWTA only* condition, and blue dotted traces for *feed-forward* condition). II. Map of the reciprocal synaptic connections between R and X, Y and Z, forming the mean function $Z = \text{floor}[(X+Y)/2]$. The IF2D chip output (winning neuron) is highlighted in red.

were applied to the neurons in Z, with uniform mean firing rates equal to the average firing rates of neurons in the base populations X and Y. Note that the input to Z did not encode the expected result. We applied the external stimulation to all three populations simultaneously for 1000 ms; spike events were recorded from the stimulus onset to 400 ms after the end of external stimulus.

B. Cue Integration

We applied a wide input Gaussian profile of firing rates (75 Hz maximum firing rate, $\sigma = 3.5$) to all three populations X, Y and Z. Inputs to X and Y were centered around random positions, chosen at least 6 neurons away from the edges to avoid boundary effects. The input to Z was chosen to be compatible with the inputs to X and Y, according to the mean function. We applied the external stimulus to the three populations simultaneously for 1000 ms, and recorded their activities for 1400 ms. As expected, the network sWTA properties sharpened the activity profiles. To quantify the sharpening effect, we used the data collected from 700 ms to 1000 ms after the stimulus onset, ensuring adequate time for R to choose a winner.

IV. Experimental Results

A. Function Approximation

The network was able to correctly calculate the mean of the two inputs in most of the trials. The activity on Z converged to within two positions of the correct neuron in 77% of trials. The network converged to within three positions in 83% of trials. The final position was read out after removing the input stimulus, from the location of the most active neuron; if several neurons shared the highest firing rate, their positions were averaged. These results are summarized in Table I. The activities measured from all three populations in a single trial are shown in Fig. 2a. In addition to the fully connected network output, we plot the neuron activities in the *feed-forward* and *sWTA only* conditions, for comparison. The raw data showing the spiking activity on population Z for a different trial is shown in Fig. 3. The most common source of error is due to R failing to activate the cell with the highest input. This is the case in almost all of the trials where activity on Z differed by more than 2 positions from the mean. When an incorrect solution was chosen on R, it tended to impose that solution on the base variables X and Y. Thus the error in Z was highly correlated with error in populations X and Y ($\rho = 0.96$).

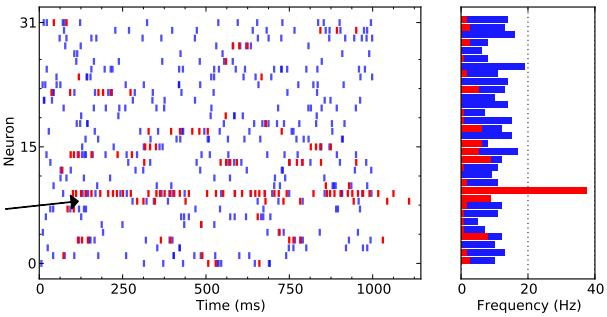

Fig. 3: Spiking activity of population Z during a function approximation trial. External input spikes are plotted in blue while the output firing activity is plotted in red. Maximum inputs to X and Y were at positions 4 and 13. Population R chose a winner at position $[3, 13]$ (corresponding to a mean of 8) at 117 ms, indicated by the arrow. Maximum activity in Z occurred at position 9. Activity continued after the end of external stimulus on positions 8 and 9.

B. Cue Integration

The inclusion of the intermediate population R in the network significantly sharpened the activity profiles, as predicted by theory [2]. We measured a 27% increase in sharpening by comparing the full versus the *sWTA only* conditions, based

TABLE I: TRIALS ENDING IN CORRECT POSITION

Distance from Correct Position	Function Approximation		Cue Integration	
	X,Y	Z	X,Y	Z
≤ 1	75%	**63%**	58%	87%
≤ 2	83%	**77%**	87%	100%
≤ 3	90%	**83%**	97%	100%

TABLE II: INPUT AND NETWORK RESPONSE WIDTHS

Network	Function Approximation	Cue Integration
	X,Y	X,Y,Z
Input	1.75	**3.50**
Feed-forward	1.48 ± 0.01	$\mathbf{2.92 \pm 0.01}$
sWTA only	0.97 ± 0.02	$\mathbf{1.60 \pm 0.03}$
Relational	1.03 ± 0.10	$\mathbf{1.17 \pm 0.04}$

on the standard deviation of the spiking rates across the population. These results are summarized in Table II; a single trial is shown in Fig. 2b. In particular, sharpening occurred in 92% of the 90 results collected (30 trials, 3 populations).

Overall, the positions of convergence were also accurate. Activity on X and Y converged to within two positions of the correct neuron in 87% of the results, and activity on Z converged to within two positions in all trials (see Table I).

In the 8% of results where the activity profile was not sharpened by the addition of R, the activity also tended to converge to the wrong position. These results were all from populations X and Y. Analysis of the associated trials showed that, in most cases, R chose a suboptimal solution that was close to being consistent with Z and one of X or Y, but was less consistent with the other base variable. The activity on the inconsistent variable was then pulled towards the solution chosen by R, producing a wide response in the time window. This contributed both to increased distance error and decreased sharpening in X and Y.

V. DISCUSSION

Our results demonstrate that neuromorphic WTA networks can be used to implement neural systems that constrain the values of multiple cues based on user-defined relationships. In the system we developed, spiking inputs to two neural populations shaped the activity of a third population according to the function defining their relationship (function approximation). We also presented cue integration experiments where broad input profiles to three populations recurrently sharpened each other to produce a consistent output determined by a user-specified relation. Unlike previous software implementations with similar capabilities, our network used a 2-D neural array that allowed only one active neuron at any point in time (hard WTA behaviour).

These promising results demonstrate that it is possible to construct a complex network with many variables and relations, by expanding the setup with additional 1-D and 2-D populations of VLSI spiking neurons. The computations our network can achieve can be used to carry out a variety of tasks, such as sensorimotor transformation, multimodal sensory integration and feature extraction, and is therefore a step towards the configuration of generic processing in neuromorphic systems. This class of neural architectures and the computations they can achieve are optimal in the Bayesian sense [3], suggesting that the system presented here can offer an efficient technology for building probabilistic models, such as Bayesian networks or Factor Graphs. Such systems can have practical applicability in robotic systems which make use of event-based neuromorphic sensors [16]. By combining the input from several sensors, beliefs about the environment can be integrated and refined, providing a general framework for multimodal cue combination in neuromorphic systems.

ACKNOWLEDGEMENT

This work was supported by the EU ERC Grant "neuroP" (257219), the EU ICT Grant "SCANDLE" (231168), by the Swiss SNF Grant "nAttention" (121713), and by the Excellence Cluster 227 (CITEC, Bielefeld University) The WTA2D chip was designed in collaboration with C Bartolozzi. We thank D. Fasnacht for the design of the AER mapper and the AER monitor/sequencer boards.

REFERENCES

[1] E. Salinas and T. Sejnowski, "Correlated neuronal activity and the flow of neural information," *Nature Reviews Neuroscience*, vol. 2, pp. 539–550, 2001.

[2] S. Deneve, P. Latham, and A. Pouget, "Efficient computation and cue integration with noisy population codes," *Nature Neuroscience*, vol. 4, no. 8, pp. 826–831, 2001.

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

[4] M. Cook, F. Jug, C. Krautz, and A. Steger, "Unsupervised learning of relations," *Artificial Neural Networks–ICANN 2010*, pp. 164–173, 2010.

[5] C. Mead, *Analog VLSI and Neural Systems*. Reading, MA: Addison-Wesley, 1989.

[6] A. L. Yuille and D. Geiger, *Winner-Take-All Networks*. The MIT Press, Cambridge, Massachussets, 2003, ch. Part III: Articles, pp. 1228–1231.

[7] R. Hahnloser, R. Sarpeshkar, M. Mahowald, R. Douglas, and S. Seung, "Digital selection and analog amplification co-exist in an electronic circuit inspired by neocortex," *Nature*, vol. 405, no. 6789, pp. 947–951, 2000.

[8] T. Binzegger, R. Douglas, and K. Martin, "A quantitative map of the circuit of cat primary visual cortex," *J. Neurosci.*, vol. 24, no. 39, pp. 8441–53, 2004.

[9] R. Douglas and K. Martin, "Neural circuits of the neocortex," *Annual Review of Neuroscience*, vol. 27, pp. 419–51, 2004.

[10] C. Eliasmith, "A unified approach to building and controlling spiking attractor networks," *Neural Computation*, vol. 17, no. 6, pp. 1276–1314, 2005.

[11] U. Rutishauser, R. Douglas, and J. Slotine, "Collective stability of networks of winner-take-all circuits," *Neural Computation*, vol. 23, no. 3, pp. 735–773, 2011.

[12] 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.

[13] D. Fasnacht and G. Indiveri, "A PCI based high-fanout AER mapper with 2 GiB RAM look-up table, 0.8 μs latency and 66 mhz output event-rate," in *Conference on Information Sciences and Systems, CISS 2011*, Johns Hopkins University, March 2011, pp. 1–6.

[14] C. Bartolozzi and G. Indiveri, "Selective attention in multi-chip address-event systems," *Sensors*, vol. 9, no. 7, pp. 5076–5098, 2009.

[15] E. Neftci and G. Indiveri, "A device mismatch reduction method for VLSI spiking neural networks," in *Biomedical Circuits and Systems Conference BIOCAS 2010*. IEEE, 2010, pp. 262–265.

[16] S.-C. Liu and T. Delbruck, "Neuromorphic sensory systems," *Current Opinion in Neurobiology*, vol. 20, no. 3, pp. 288–295, 2010.