Self-Evolving Neuromorphic Networks with Single-Electron Switching Latches as Synaptic Nodes

Simon Fölling, Özgür Türel, and Konstantin Likharev*

***Abstract* – 2D square arrays of single-electron latching switches may serve as a synaptic basis for extremely dense self-evolving BiWAS (binary weight, analog signal) neuromorphic networks. Two promising architectures of such networks have been preliminary studied: a "free-growing" network in which the shape of axonic and dendritic trees may be very complex, and a "randomized distributed crossbar" network in which axons and dendrites are implemented as straight wire segments. The latter network scales much better, but the former one may be more adequate for input parts of bio-scale networks.**

1 Introduction

Future ULSI implementation of bio-scale neural networks requires nanoscale synaptic devices. Indeed, in order to place a network with $N = 10^{10}$ neurons with connectivity $M = 10^4$, on a 10×10 cm^2 chip, the synapse should fit onto a 10×10 nm^2 area. Such density may be achieved using single-electron devices [1]. Earlier work on such devices in the neural network context [2] was focused on the implementation of neural cell bodies. On the contrary, we believe that these (much less numerous) components may be left for the mainstream CMOS technology, thus circumventing such problems of single-electron devices as low transconductance and random background charge effects [1].

2 Synapse

We have designed several single-electron devices which may serve as a BiWAS (binary-weight, analog-signal) synaptic nodes bridging two nanowires. The simplest of these devices (Fig. 1) consists of three small conducting islands. Island 1, together with input and output wires, forms a single-electron transistor [3-5]. Islands 2 and 3 form a single-electron trap [6], with island 3 capacitively coupled to the transistor island. If voltage $V = V_S - V_D$ between the wires is low, the trap in equilibrium has no trapped electrons ($n = 0$). As a result, the transistor remains in the Coulomb blockade state, and input and output wires are essentially disconnected.

If V is increased beyond a certain threshold V_{inj} (which should be lower than the Coulomb blockade threshold voltage V_t of the transistor), one electron is injected into the trap ($n \to 1$). In this charge state the Coulomb blockade in the transistor is suppressed, keeping the wires connected at any V. However, if the node activity (voltage V) is low for a long time, the electron tunnels out of the trap ($n \to 0$) and the transistor closes, disconnecting the wires.

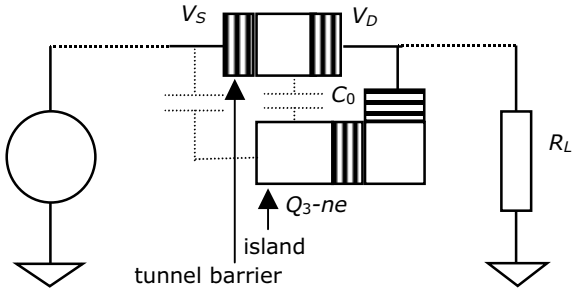

Figure 1: A single-electron latching switch playing the role of a binary weight synapse.

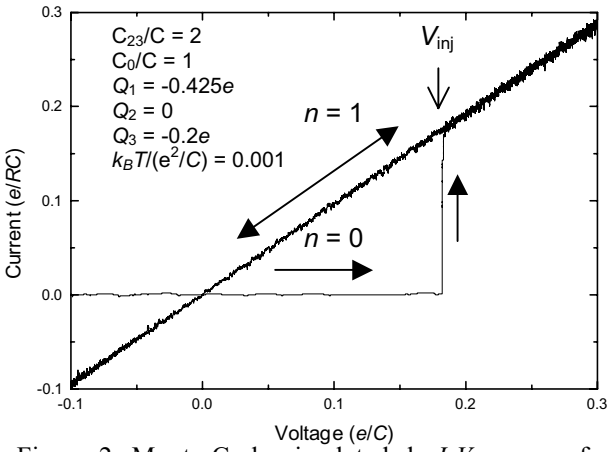

Figure 2: Monte-Carlo simulated dc I-V curves of the latching switch shown in Fig. 1. Q_i are island background charges; C and R are, respectively, the capacitance and resistance of most tunnel junctions of the circuit. (The only exception is indicated on the inset.)

The results shown in Fig. 2 are typical for the latching switch with a perfect set of island background charges Q_i. (They may be fixed by gating through small capacitances [1] not shown in Fig. 1.) If the charges deviate from these values (say due to randomly located charged impurities [1]), the Coulomb blockade of the transistor is not completely

*. State University of New York, Stony Brook, NY 11794-3800, U.S.A. E-mail klikharev@notes.cc.sunysb.edu, phone +1-631-632-8159, fax +1-631-632-8774.

suppressed even in the "open" state of the node (n = 1). For massively parallel circuits like neuromorphic networks, this imperfection should be, however, quite tolerable.

3 Free Growing Network

We have carried out preliminary studies of two promising architectures of adaptive ("plastic") neuromorphic networks based on 2D square arrays of the single-electron switches.

In the first architecture, axonic and dendritic trees grow spontaneously on the array of 8×8 switching nodes (Fig. 3), each consisting of 24 binary switches, some of them similar to that shown in Fig. 1 and some slightly more complex. These switches may connect each input wire to its straight continuation and (with a lower probability) to two other output wires forming angles ±45° with the incoming wire. Probability of each connection depends on the signal amplitude and hence on the previous length of the axon. Figure 4 shows a typical picture of axonic (red) and dendritic (blue) trees growing from a neural cell body.

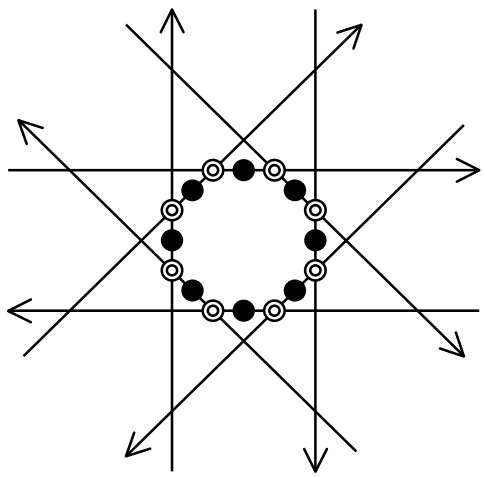

Figure 3: An 8×8 switching node for self-growing networks. Each of 8 input wires may be connected to one output line via one "propagating" switch (black point, Fig. 1) and two more output wires via "branching" switches (overlapped open points).

If complemented by a special simple circuit (essentially a double-directional diode) which mutually grounds axonic (positive) and dendritic (negative) voltages in the node where an axon and a dendrite meet, this model forms patterns (see, e.g., Fig. 5) which are strikingly similar to those observed in biological neural networks.

This behavior seems very promising, but unfortunately we have found that the free growing networks do not scale well with the circuit connectivity M (the average number of neural cells connected to a given cell). Specifically, if F is a minimum feature size available in a given fabrication technology, the average distance x between two neural cell bodies is as high as

$$x \sim 25M^{3/2}F. \qquad (1)$$

Simple estimates show that even at F = 1 nm (this technology level allows room temperature operation of single-electron devices [1]), at the connectivity typical for the cerebral cortex ($M \sim 10^4$), density of the free growing networks can hardly exceed a few neurons per cm^2, the level evidently too low to model brain-scale systems. This is the price we pay for the fact that a huge number of axonic and dendritic trees may lead to the same set of synaptic weights, i.e., to virtually the same network behavior. We still believe such networks, with much lower M, may be useful in input signal processing sub-systems, e.g., artificial retinas.

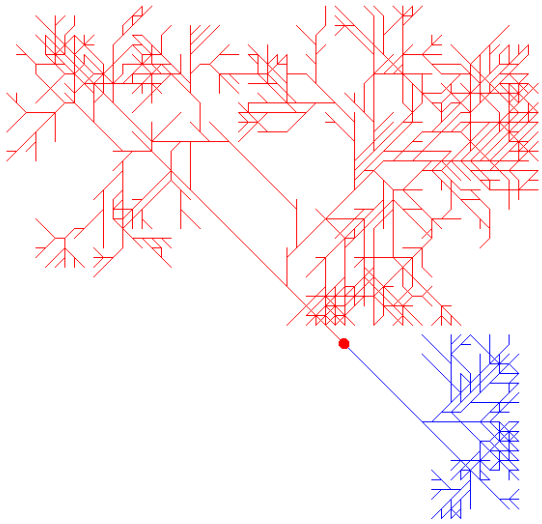

Figure 4: A typical pattern of the initial growth of an axonic (red) and dendritic (blue) trees from a neural cell body (red dot), stimulated by its activity. All the plane is filled with a 2D array of the nodes shown in Fig. 3, but only activated (voltage carrying) links are shown.

4 RandBar

Much better scaling may be achieved in another, "randomized distributed crossbar" (RandBar) architecture shown in Fig. 6. In this network, each cell is hard-wired to a limited subset of other cells, with the binary synaptic weights controlling which of these connections are currently active. This network actively uses virtually the whole chip area.

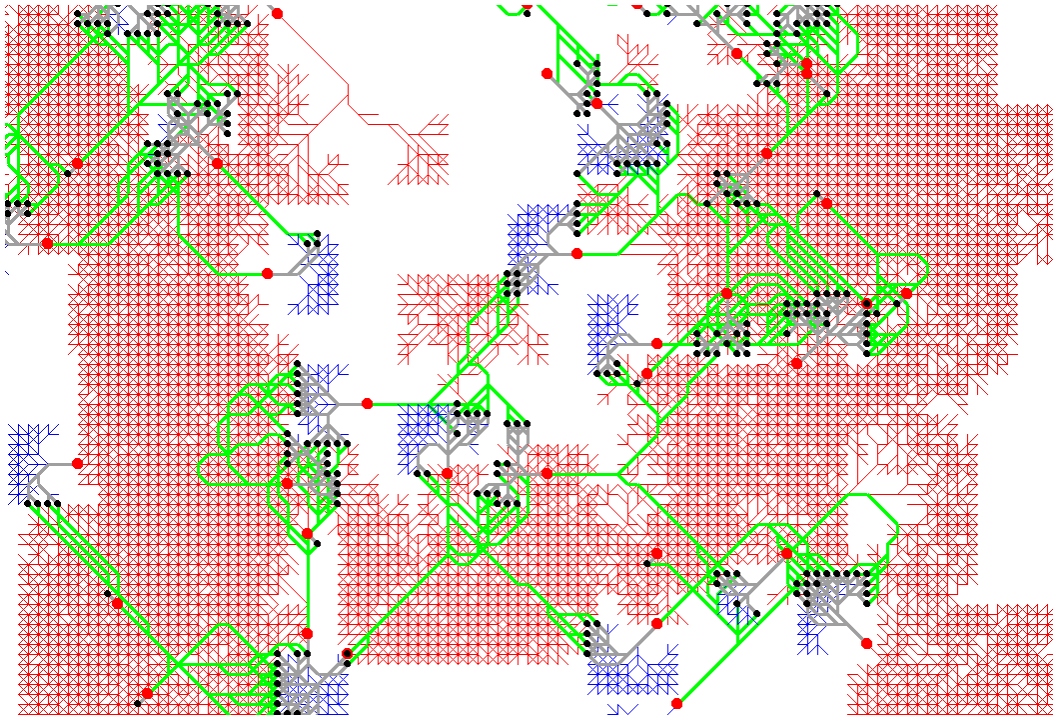

Figure 5: A typical network, free growing from several active, randomly located, neural cell bodies (red dots). As in Fig. 4, red lines are growing axons and blue lines are growing dendrites, while gray and green lines show "frozen" axons and dendrites, respectively. The "freeze" takes place when an axon and a dendrite meet and form a synapse (black dot). The total number of synapses formed between an axonic tree of one neuron and a dendritic tree of another neuron gives the corresponding synaptic weight.

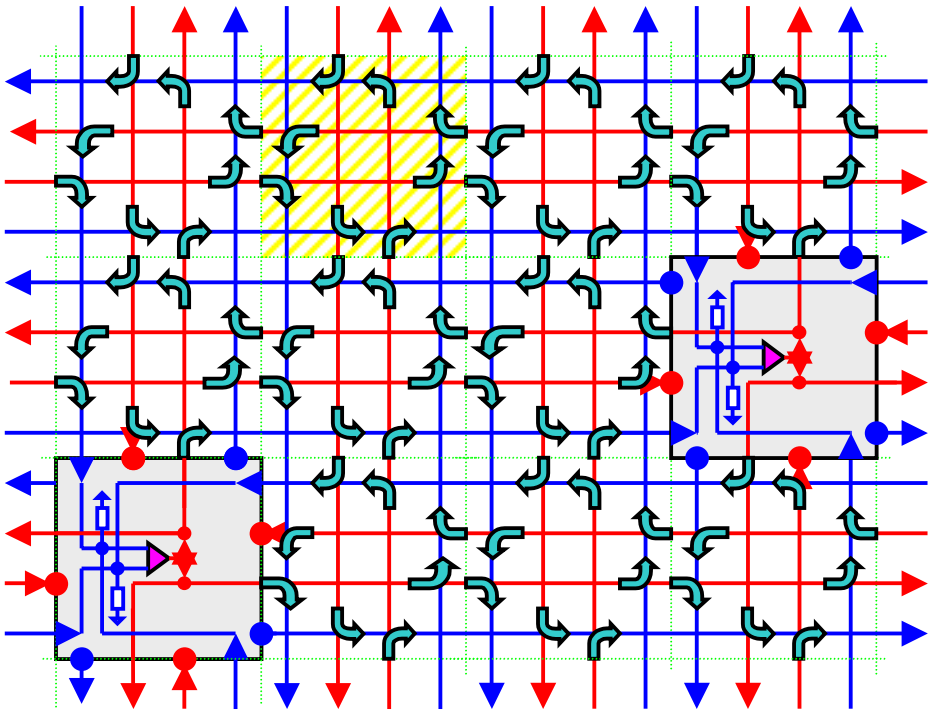

Figure 6: "RandBar". Each neural cell body (gray) sends its output signal into 4 axonic lines (red) and receives signals from 4 dendritic lines. The lines are connected by single-electron switches (shown by the curly green arrows), similar to one shown in Fig. 1. The yellow square indicates the basic cell of the switching node/wiring structure.

As a result, scaling of this network is much more favorable:

$$x \sim 15\, M^{1/2}\, F. \qquad (2)$$

For the example given above ($M = 10^4$, $F = 1$ nm), this scaling gives a density estimate almost as high as 10^8 neurons per cm^2, implying that brain-scale systems could be implemented on the chip area of the order of 100 cm^2 – the size which the electronics industry plans to reach in just about 10 years [7].

Estimated speed scaling of this network is also very impressive. Estimates show that the characteristic time of whole network "settling" (apparently, equivalent to one learning cycle, of the order of a few seconds for biological neural networks) should be as short as

$$\tau_s \sim \tau_0\, N/M, \qquad (3)$$

where τ_0 is the time of information propagation by one network layer, physically dominated by charging of dendrite wire capacitances through a relatively high resistance $R >> R_Q \sim h/e^2 \sim 10^4$ Ohms [1] of opened single-electron transistors. For $M = 10^4$ and $F \sim 1$ nm, simple estimates give $\tau_0 \sim 0.1$ ns. (This high speed is a result of absence of very long interconnects in the quasi-local architecture of the network.) As a result, even for a very large network ($N = 10^{10}$) we get τ_s as small as 10^{-5} s, about 6 orders of magnitude faster than the biological evolution. In order to comprehend the significance of this estimate, it implies that one evolution iteration (equivalent to one biological generation, of the order of 10 years for mammals) may be achieved in *just a few seconds*!

It is also instructive to compare the speed of this "hardware" implementation of self-evolving neuromorphic networks with that of their "software" implementation on general-purpose computers (the presently dominating approach to neural networks). One τ_s corresponds to approximately N^2, i.e. to about 10^{20} updates for the parameters quoted above. This means that the most powerful existing parallel supercomputer, with the peak performance of a few teraflops, would spend at least 100 days for just one settling, and (impractical) 10^8 years for just one evolution cycle. This comparison gives a spectacular example of the possible information processing power of the "ultra-parallel" self-evolving integrated circuits.

Preliminary estimates of another factor, power dissipation (which imposes very strict restrictions on the performance of the traditional CMOS VLSI circuits [7]), also give acceptable results: they show that the main contribution to the dissipation will be given by CMOS circuits used for cell bodies. For the $F = 1$ nm technology level, this power should be of the order of 100 W/cm^2 – a little bit high to be pleasant, but quite manageable.

A possible drawback of the network architecture shown in Fig. 6 is the binary character of the synaptic weights $w_{ij} = \{0,1\}$. We hope that effects of this discreteness will be partly compensated by the randomness of each connection (which is due to both the random location of cell bodies and the background charge randomness). However, only detailed simulation of such networks may show whether this limitation imposes undesirable restrictions on unsupervised learning.

5 Conclusions

We believe that room-temperature bio-scale systems may be implemented on single chips with 1-nm-scale nanofabrication technologies (presently under active development in several laboratories). Speed scaling of these network is also very impressive: they should process information and self-evolve about 6 orders of magnitude faster than biological systems.

At the meeting, we intend to present detailed results of synaptic node modeling, and of simulated behavior of reasonably large network fragments (with up to 10^6 synapses), both in the autonomous mode and under globally supervised learning.

References

[1] See, e.g., K. Likharev, "Single-Electron Devices and Their Applications" Proc. of IEEE, vol. 87, 1999, pp. 606-632.

[2] M. Akazawa and Y. Ameniya, "Boltzmann Machine Neuron Circuit Using Single-Electron Tunneling", Appl. Phys. Lett., vol. 70, 1997, pp. 670-673.

[3] D. Averin and K. Likharev, "Coulomb blockade of tunneling", J. Low Temp. Phys., vol. 62, 1986, pp. 345-372.

[4] K. Likharev, "Single-Electron Transistors: Electrostatic Analogs of the DC SQUIDs", IEEE Trans. Magn., vol. 23, 1987, pp. 1142-1145.

[5] T. Fulton and G. Dolan, "Observation of single-electron charging effects in small tunnel junctions", Phys. Rev. Lett., vol. 59, 1989, pp. 109-112.

[6] P. D. Dresselhaus, L. Ji, S. Han, J. Lukens, and K. Likharev, "Measurement of single-electron lifetimes in a multijunction trap", Phys. Rev. Lett. vol. 72, 1994, pp. 3226-3229.

[7] "International Technology Roadmap for Semiconductors", see http://http://public.itrs.net/.