Published in: *Proc. IJCNN'01*
(Int. Neural Network Society, Mount Royal, NY), pp. 216-221

Single-Electron Latching Switches as Nanoscale Synapses

Simon Fölling, Özgür Türel, and Konstantin Likharev

State University of New York at Stony Brook
Stony Brook, NY 11794-3800, U.S.A.
klikharev@notes.cc.sunysb.edu

Abstract

Nanoscale latching switches based on controllable single-electron transfer and trapping may serve as a synaptic basis for extremely dense and fast self-evolving BiWAS (binary weight, analog signal) neural networks. We have designed and simulated two devices of this type, a "propagating" switch and a "branching" switch, as well as multi-entry switching nodes based on their combination. We have also carried out a preliminary study of two architectures of neural networks based on 2D arrays of the switching nodes: 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 very large scale networks.

1 Introduction

Hardware implementation of neural networks comparable in complexity with the cerebral cortex requires nanoscale synaptic devices. Indeed, in order to place a network with $N = 10^{10}$ neurons with the average 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 which allow controllable transfer and trapping of single electrons in system of small conductors ("islands") separated by tunnel barriers.

Physics of single-electron devices [1] is based on the so-called "Coulomb blockade" effect: If the size of an island and hence its capacitance C are so small that the electrostatic energy $e^2/2C$ of its charging by a single electron is well above the thermal fluctuation energy scale k_BT, this charging may prevent ("block") transfer of other electrons into this island and also alter electron transport in the adjacent islands.

The enormous advantage of single-electron devices, in comparison with the mainstream, "CMOS" technology of electronic circuits, is the possibility of their scaling down to virtually single-atom size (islands smaller than 1 nm), and hence reaching unparalleled digital device densities up to 10^{14} cm^{-2}. Another important feature of these devices is the virtual independence of their operation on the parameters of used materials, giving in particular a strong hope for their implementation using molecular self-assembly.

On the other hand, digital single-electron devices have substantial drawbacks, including notably low transconductance and the infamous random background charge effect (considerable spread of single-electron device switching thresholds upon the effect of randomly located charged impurities). The latter effect may be overcome in single-electron memories [1-3], but makes it virtually impossible to use single-electron devices in usual logic circuits. However, neural networks with their high parallelism looks like a perfect application for single-electronics.

Earlier work on single-electron devices in the neural network context was focused on the implementation of neural cell bodies – see, e.g., Ref. 4. On the contrary, we believe that these (considerably less numerous) components may be left for the CMOS technology, thus circumventing the problems of single-electron devices, which were mentioned above. On the other hand, these devices are virtually perfect for implementation of simple BiWAS (binary-weight, analog-signal) synapses.

2 Single-electron Synapses

Figure 1 shows the simplest single-electron latching switch which may serve as a synaptic node bridging two nanowires. The device consists of three small islands connected by four tunnel junctions. Island 1, together with input and output wires, forms a single-electron transistor [5-7]. Conductance of this device for small applied source-drain voltages may be very low (the "Coulomb blockade state") unless the blockade is lifted by electric field applied by a special gate electrode [1]. In our case the role of the gate is played by another island (3) which also forms (together with island 2) another device, a single-electron trap [8].

If the source-drain voltage $V = V_S - V_D$ between the wires is low, the trap in equilibrium has no extra electrons and its total electric charge is zero. 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. In this charge state the Coulomb blockade in the transistor is lifted, keeping the wires connected at any V. However, if the node activity (voltage V) is low for a long time, unavoidable thermal fluctuations eventually kick the trapped electron out of the trap and the transistor closes, disconnecting the wires.

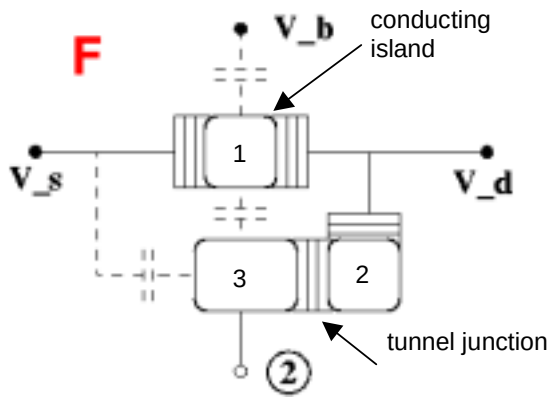


Figure 1: The simplest single-electron latching switch which may be used as a "forwarding" BiWAS synapse.

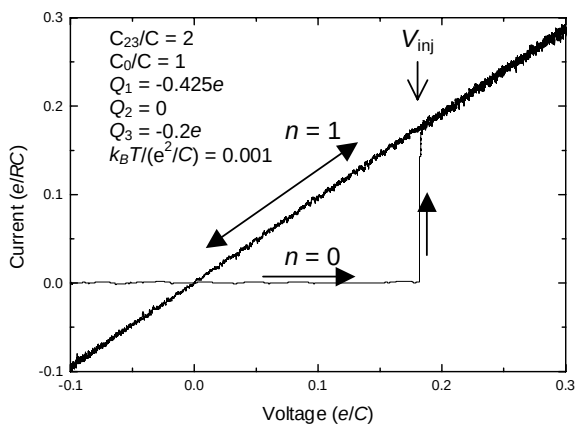


Figure 2: Monte-Carlo simulated dc I-V curves of the forwarding node shown in Fig. 1. Q_i are island background charges; C and R are, respectively, the capacitance and resistance of most tunnel junctions of the circuit. (The only exception is indicated on the inset.) The simulation was based on the "orthodox" theory of single-electron tunneling which is quantitatively correct for systems with islands of not very small size (≥ 1 nm).

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 a proper adjustment of additional voltage V_b – see Fig. 1.) If the charges deviate from these values (e.g., due to randomly located charged impurities [1]), the Coulomb blockade of the transistor is not completely suppressed even in the "open" state of the node. For massively parallel circuits like neural networks, this imperfection should be, however, quite tolerable.

Figure 3 shows a slightly more complex latching switch. An additional island 4 forms a "single-electron box" [9]. In contrast to the trap, the number of electrons in the box is a unique function of the applied voltage, in this particular case V. Namely, if V exceeds a certain value V_{box}, an extra electron tunnels into island 4. Due to the Coulomb repulsion, this injection makes passage of an electron to trap 3 impossible. (Essentially, island 2 now plays a role of an additional single-electron transistor connecting island 3 to the common drain.) As a result, the transistor with island 1 remains closed for both very low and very high voltages, and opens only in an intermediate voltage range.

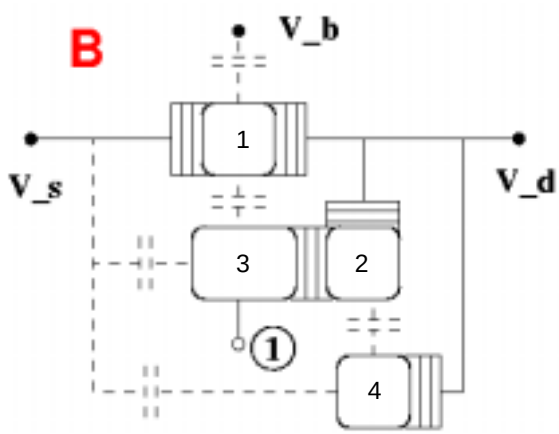


Figure 3: The "branching" latching switch.

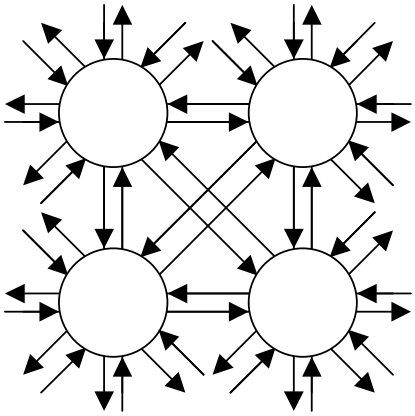
3 Free Growing Neural Network

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

In the first architecture, axonic and dendritic trees grow spontaneously on a 2D array of 8×8 nodes (Fig. 4a). Each of 8 output signals may be contributed by 3 incoming wires: one along the output direction ("forwarding") and two other input wires forming angles ±45° with the outgoing wire. This is achieved using a composite switch (Fig. 4b) consisting of one forwarding latch (Fig. 1) and two branching latches (Fig. 3). Capacitive coupling of these devices ensures their mutual blocking so that only one input is actually connected to each output. Thus the switching node as a whole (Fig. 4a), consisting of 8 composite latches, allows up to 8 input signals to be forwarded or branched without mutual interference. Due to the properties of the F and B circuits (Fig. 1 and 3), probability of each connection depends on the signal amplitude and hence on the previous length of the axon/dendrite. At short

distances from the signal source (neural cell body) these properties favor forwarding (straight "growth" of the wires carrying nonvanishing voltage, called active links), while at larger distances, branching becomes almost equally probable.

(a)



(b)

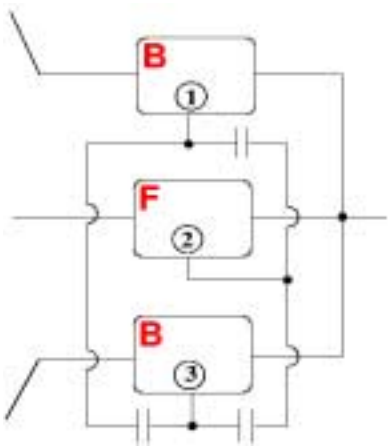
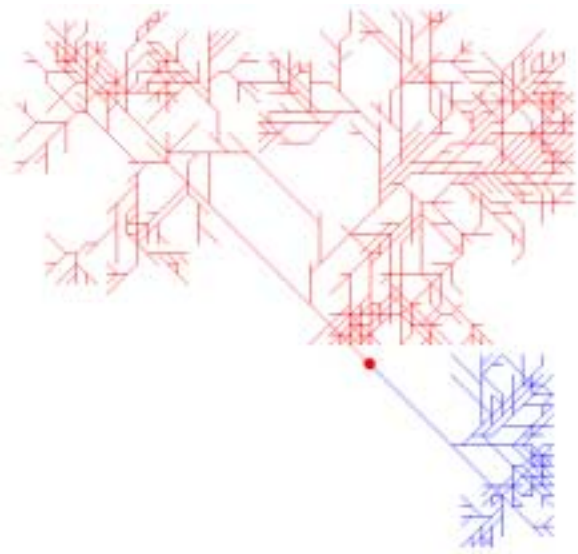


Figure 4: (a) A 2D array network of 8×8 switching nodes for self-growing networks and (b) a latching switch composed of elementary single-electron latches shown in Fig. 1 and 3.

We have Monte-Carlo-modeled the initial evolution of self-growing networks formed under the effect of output signals of neural cell bodies which are randomly placed on two similar, overlapping 2D switching node arrays, one for dendrites and one for axons. (Parameters of dendritic switches were slightly different, providing more branching for dendrites than for axons.) Figure 5 shows a typical picture of axonic (red) and dendritic (blue) trees growing from a neural cell body.

If complemented by a special simple synapse circuit (essentially a double-directional diode) which mutually shortens axonic (positive) and dendritic (negative) voltages in the node where an axon and a dendrite meet, this model grows in patterns (see, e.g., Fig. 6) which are strikingly similar to those observed in biological neuron 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 in this architecture is as high as



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

Figure 5: 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 8×8 switching nodes shown in Fig. 4a, but only activated (voltage carrying) links are shown.

Simple calculation shows 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 in this architecture a huge number of axonic and dendritic trees may lead to the same set of synaptic weights, i.e., eventually to 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, but for the core signal processing a less redundant architecture is necessary.

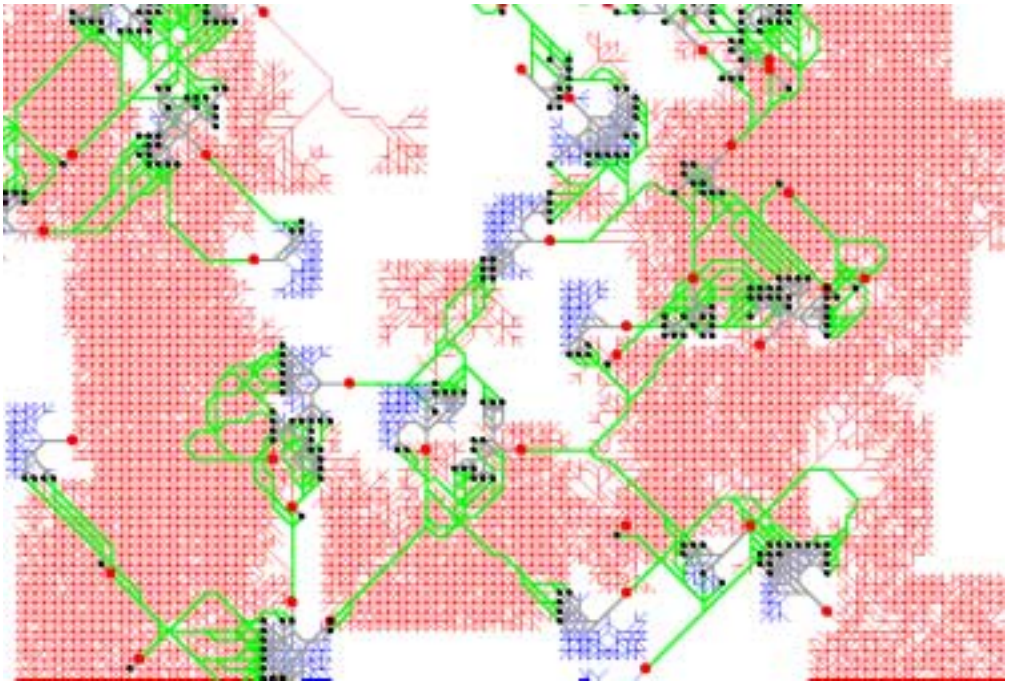


Figure 6: A typical pattern of the initial growth from several active, randomly located, neural cell bodies (red dots). As in Fig. 5, red lines are growing axons and blue lines are growing dendrites, while gray and green lines show "frozen" axons and dendrites, respectively. The "freeze" takes place 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. (At this stage of modeling, neural cell signals were considered fixed, i.e., no actual signal dynamics was traced.)

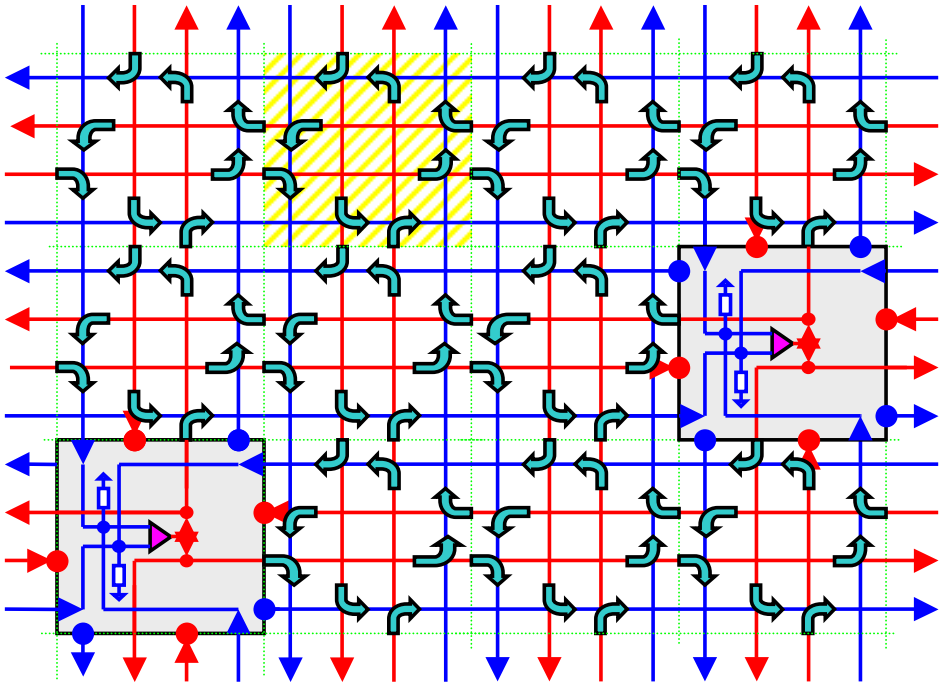


Figure 7: "RandBar". Each neural cell body (gray square) sends its output signal into 4 axonic lines (red) and receives signals from 4 dendritic lines. The lines are connected by simple single-electron switches (shown by the curly green arrows) whose structure is shown in Fig. 1 [10]. Each switch plays the role of a BiWAS synapse. The yellow square and dashed green lines are just guides for the eye, indicating the basic cells of the switching node/wiring array.

4 RandBar

Much better scaling may be achieved in another, "randomized distributed crossbar" (RandBar) architecture (Fig. 7). As before, cell bodies are embedded randomly into a 2D array of single-electron latching switches, but now every switch plays the role of a BiWAS synapse. As a result, 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; as a result, scaling improves dramatically:

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

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 a chip area of the order of 100 cm^2 – the size which the electronics industry plans to reach in just about 10 years [10].

Estimated speed scaling of this network is also very impressive. Time scale τ_0 of signal propagation by one network layer is 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 open single-electron transistors. For $M = 10^4$ and $F \sim 1$ nm, simple estimates give $\tau_0 \sim 0.1$ ns. (This is a result of a small average length, about 1 micron, of interconnects in the quasi-local architecture of the network.) This speed is 6 to 7 orders of magnitude (!) higher than that of cerebral cortex cells.

Preliminary estimates of another important factor, power dissipation (which imposes very strict restrictions on the performance of the traditional CMOS VLSI circuits), 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, but still manageable.

We have carried out a preliminary Monte-Carlo modeling of dynamics of modest (N up to 10^5, NM up to 10^6) RandBar fragments, at this stage ignoring the effect of signals on the state of synaptic latches. (For the simulations shown here , 50% of the latches, at random locations, have been connected). The dynamics depend essentially on the net neural cell gain G (including signal attenuation in synapses). If the gain is less than some critical value G_t (depending on the average connectivity M), the network is "dead": signals do not change in time, though they may differ from wire to wire because of random position of open and closed synapses. If G is increased beyond G_t (≈ 0.5 for the case illustrated in Fig. 8) the network starts generating random oscillatory signals. Close to G_t these signals are typically almost sinusoidal, with a period of a few RC – see red lines in Fig. 8. At somewhat larger G, several quasi-independent oscillators of this type may drift over the network, at each particular instant being localized to a region with a size of the order of M. Finally, large gain (see black lines in Fig. 8) leads to intensive, essentially chaotic oscillations of almost the entire network.

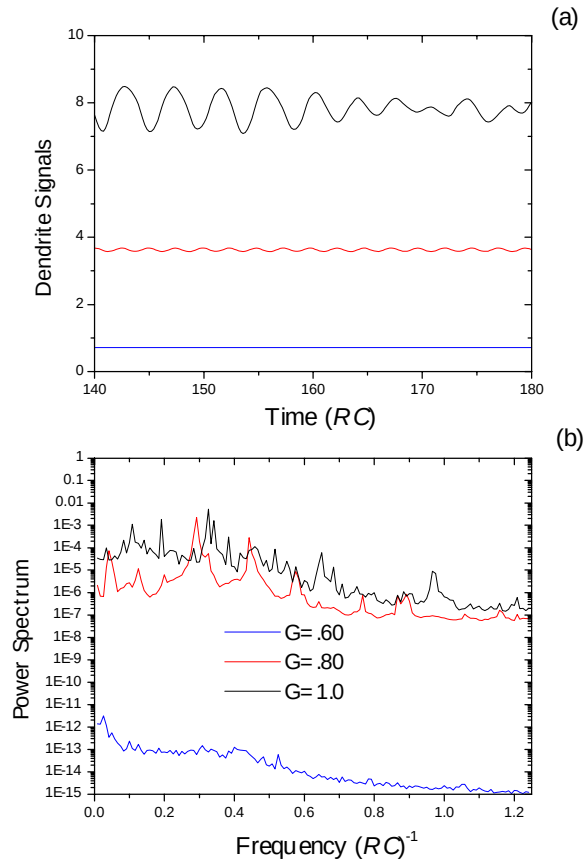


Figure 8: (a) Typical oscillograms of dendrite signals and (b) and their power spectrum, for a rectangular RandBar fragment with N = 4,800 neural cell bodies embedded randomly into an array of 100×800 switching node plaquettes. Each plaquette (like one shaded yellow in Fig. 7) consists of 8 elementary synapses similar to those shown in Fig. 1, so that the average cell connectivity M = 80,000×8/4,800 ≈ 133. Cell bodies are modeled by perceptrons, with differential inputs (see the gray shaded rectangles in Fig. 7) and the usual sigmoid saturation function. RC is the time constant of the dendrite wire recharging through a single connected synapse, G is the net gain of one network level, for a signal below the saturation threshold. Signal amplitude in (a) is normalized to the perceptron saturation level. Power spectrum in (b) is averaged over the long side of the RandBar fragment (800 outputs) and is normalized to the value which would be provided by 800 signal oscillating with the same frequency and maximum amplitude. (Due to averaging over 300 time points, in these units a set of random, unit-amplitude outputs would give a flat power spectrum level of 1/300.)

These properties of the RandBar are qualitatively understandable, but so far defy our attempts at their quantitative analysis. The main reason of this is that due to the sign symmetry of the perceptron inputs (Fig. 7) this structure (despite a long range of cell interaction) lacks a static order parameter even at large G. Statistics of random systems of this type can very rarely be traced analytically [11]. We believe, however, that this richness of behavior is promising rather than prohibiting from the point of view of information processing [12, 13]. Our plans are to carry out extensive, large-scale Monte-Carlo simulations of RandBar networks to confirm this assumption.

5 Discussion

Single-electron devices may provide the first plausible opportunity for the implementation of room-temperature systems comparable in complexity with the cerebral cortex, on single chips, using 1-nm-scale nanofabrication technologies which are presently under active development in several laboratories. This is especially true concerning the RandBar architecture which apparently allows to implement the maximum possible density of synapses per unit area.

The speed of information propagation in such systems may be some 7 orders of magnitude higher than in the biological neuron systems. If this scaling may be sustained through the whole hierarchy of neural activity including learning and self-evolution, it will imply that one iteration of evolution of these networks (apparently, functionally 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 such hardware implementation of self-evolving neural networks with that of their software implementation on general-purpose computers. In a hardware-implemented system with $N = 10^{10}$, $M = 10^4$, approximately $MN = 10^{14}$ synapse state updates could be made in ~ 0.1 ns, while the most powerful existing parallel supercomputers, with the peak performance of a few teraflops, would spend at least $10^{14} \times 100\text{ns} = 10^7\text{s} \approx 100$ days for such an update. This comparison gives an idea of the possible information processing power of the "ultra-parallel" self-evolving integrated circuits.

An evident drawback of the RandBar 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 analog nature of dendritic and axonic signals and randomness of each connection (which is due to both the random location of cell bodies and the background charge randomness). Our hopes are that our future detailed simulation of such networks will prove that this limitation still allows effective globally supervised learning. (Such learning may be provided changing not only the input signals, but also the common bias voltage and hence latching thresholds of the single-electron synapses, in accordance with the observed network behavior.)

6 Acknowledgments

Useful discussions with J. Barhen, M. Bender, T. Ishii, and J. Wells are gratefully acknowledged.

7 References

[1] See, e.g., K. Likharev, "Single-Electron Devices and Their Applications", *Proc. of IEEE*, vol. 87, 1999, pp. 606-632.
[2] K. K. Likharev and A. N. Korotkov, "Ultradense Hybrid SET/FET Dynamic RAM: Feasibility of Background-Charge-Independent Room-Temperature Single-Electron Digital Circuits", in *Proc. of 1995 ISDRS.* Charlottesville, VA: Univ. of Virginia, 1995, pp. 355-358.
[3] C.D. Chen, Y. Nakamura, and J.S. Tsai, "Aluminum single-electron nonvolatile floating gate memory cell", *Appl. Phys. Lett.*, vol. 71, 1997, pp. 2038-2040.
[4] M. Akazawa and Y. Ameniya, "Boltzmann Machine Neuron Circuit Using Single-Electron Tunneling", *Appl. Phys. Lett.*, vol. 70, 1997, pp. 670-673.
[5] D. Averin and K. Likharev, "Coulomb blockade of tunneling", *J. Low Temp. Phys.*, vol. 62, 1986, pp. 345-372.
[6] K. Likharev, "Single-Electron Transistors: Electrostatic Analogs of the DC SQUIDs", *IEEE Trans. Magn.*, vol. 23, 1987, pp. 1142-1145.
[7] T. Fulton and G. Dolan, "Observation of single-electron charging effects in small tunnel junctions", *Phys. Rev. Lett.*, vol. 59, 1989, pp. 109-112.
[8] 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.
[9] J. Lambe and R. C. Jaklevic, "Charge-quantization studies using tunnel capacitor", *Phys. Rev. Lett.*, vol. 22, 1969, pp. 1371-1375.
[10] "International Technology Roadmap for Semiconductors. 1999 Edition/2000 Update", see http://public.itrs.net/.
[11] See, e.g., J. M. Yeomans, *Statistical Mechanics of Phase Transitions*, Clarendon Press, Oxford, UK, 1992.
[12] R. Beale and T. Jackson, *Neural Computing*, Adam Hilger, Bristol, UK, 1990.
[13] S. Haykin, *Neural Networks*, Prentice Hall, Upper Saddle River, NJ, 1999.