Neuromorphic Architectures for Hybrid Nanoelectronic Circuits

Jung Hoon Lee, Xialong Ma, and Konstantin K. Likharev

Stony Brook University, Stony Brook, NY 11794-3800

Abstract

This report summarizes recent results in the development of bio-inspired "CrossNet" architectures for future hybrid semiconductor/nanodevice ("CMOL") circuits. In particular, we have shown that despite hardware-imposed limitations, a simple weight import procedure allows the CrossNets using simple two-terminal nanodevices to perform functions (such as image recognition and pattern classification) that had been earlier demonstrated in neural networks with continuous, deterministic synaptic weights. Moreover, CrossNets can also be trained to work as classifiers by the error-backpropagation method without a dedicated error propagation network. Finally, there are preliminary indications that CrossNets may be also trained by global reinforcement. Due to the unparalleled performance capability of CMOL circuits, such training may eventually lead to single-wafer systems with cerebral-cortex-scale integration, operating at a speed several orders of magnitude higher than that of their biological prototypes.

Keywords: Nanoelectronics, nanowires, CMOS, neuromorphic networks, training, synaptic weight import, error backpropagation, global reinforcement

1. Introduction: CMOS and CMOL

During the "heroic" age of bio-inspired artificial neural network, there had been a wide-spread hope (see, e.g., Ref. 1) that hardware implementation of such networks using advanced semiconductor VLSI circuits would allow to bring their performance on par with their biological prototypes. However, there are strong indications [2, 3] that the exponential, "Moore-Law" progress of the current VLSI paradigm (based on lithographic formation of CMOS circuits) will stall at the level of few billion functions per cm^2, i.e. several orders of magnitude below the areal density of synapses in the human cerebral cortex. The most fundamental reason for this "red brick wall" is that at the minimum feature size ~10 nm, the sensitivity of parameters of semiconductor field-effect transistors to inevitable fabrication spreads grows exponentially. As a result, the gate length should be controlled with an angstrom-scale accuracy, far beyond the long-term expectations of the semiconductor industry [2].

This is why there is a rapidly growing consensus that the impending crisis of the Moore Law may only be resolved by a radical paradigm shift from the lithographic definition of electron devices to their "bottom-up" fabrication. In the latter approach, the smallest devices should be formed in some special way (for example, synthesized chemically), ensuring their fundamental reproducibility. An example of such unit is a specially designed and synthesized molecule comprising of a few tens or hundreds of atoms. The first molecular devices of this kind have been already demonstrated experimentally.

Since nanodevices have limited functionality [3], the only plausible way toward high-performance nanoelectronic circuits is to integrate such devices, together with connecting nanowires, with CMOS chips whose transistors would provide the circuit with the necessary additional functionality, in particular high voltage gain. In particular, we have suggested a specific type of CMOS/nanodevice hybrids, called "CMOL" (standing for CMOS/MOLecular circuits) [3, 4] that provides efficient interfacing between semiconductor transistors and nanodevices.

Figure 1 shows the CMOL circuit structure. Two-terminal nanodevices are formed at each crosspoint of a crossbar array, consisting of two levels of parallel nanowires. The nanodevice/CMOS interface is provided by pins that are distributed all over the circuit area, on the top of the CMOS stack. (The technology necessary for fabrication of tips with nanometer-scale points has been already developed in the context of field-emission arrays.) Pins of each type (reaching to the lower and upper nanowire level) are arranged into a square array with side $2\beta F_{CMOS}$, where F_{CMOS} is the half-pitch of the CMOS subsystem, and β is a dimensionless factor larger than 1, that depends on the CMOS cell complexity.

The most nontrivial point of the CMOL concept is that the nanowire crossbar is turned by angle $\alpha = \arcsin(F_{nano}/\beta F_{CMOS})$ relative to the CMOS pin array, where F_{nano} is the nanowiring half-pitch. This "incline" ensures that a shift by one each nanowire pitch $2F_{nano}$

corresponds to reaching the next interface pin. As Fig. 1b illustrates, this approach allows a unique access of the CMOS subsystem to any nanodevice, even if $F_{nano} << F_{CMOS}$. Moreover, the CMOL topology ensures a modest circuit yield loss even when the nanowire fabrication technique lacks precise alignment (as it is the case for such prospective technology as nanoimprint).

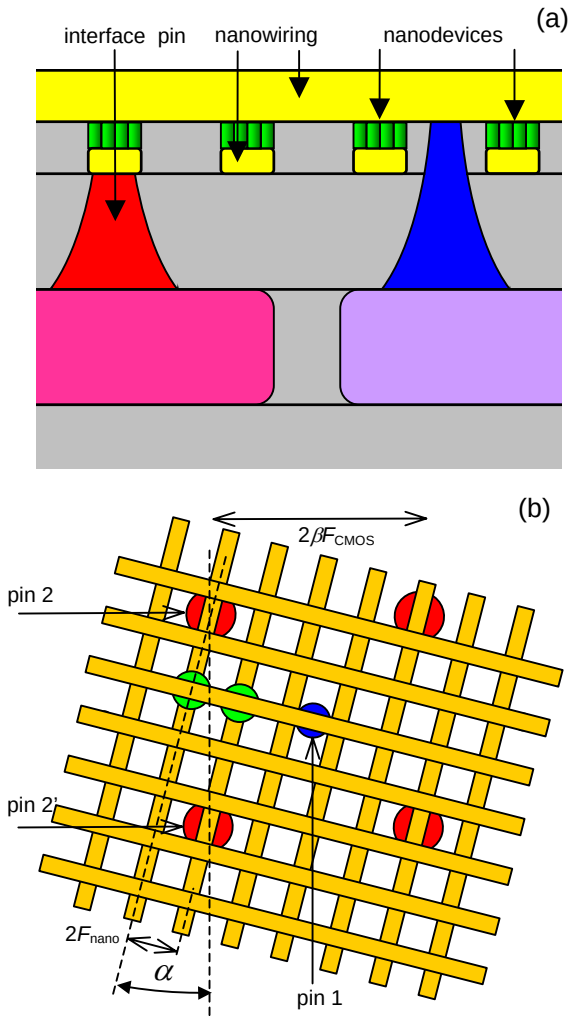

Fig. 1: Low-level structure of the generic CMOL circuit: (a) schematic side view and (b) top view showing several adjacent pins. One can see that any nanodevice may be addressed via the appropriate pin pair (e.g, pins 1 and 2 for the left of the two shown devices, and pins 1 and 2' for the right device). On panel (b), only two nanodevices are shown, while in reality, similar nanodevices are formed at all nanowire crosspoints.

If the nanodevices have a sharp current threshold, like the usual diodes, such access allow to test each of them individually. Moreover, if the device may be switched between two internal states (Fig. 2a) as, e.g., the single-electron latching switches (Fig. 2b [3-5]), each device may be turned into the desirable (ON or OFF) state by applying voltages $\pm V_W$ to the selected nanowires, so that voltage $V = \pm 2V_W$ applied to the selected nanodevice exceeds the corresponding switching threshold, while half-selected devices (with $V = \pm V_W$) are not disturbed.

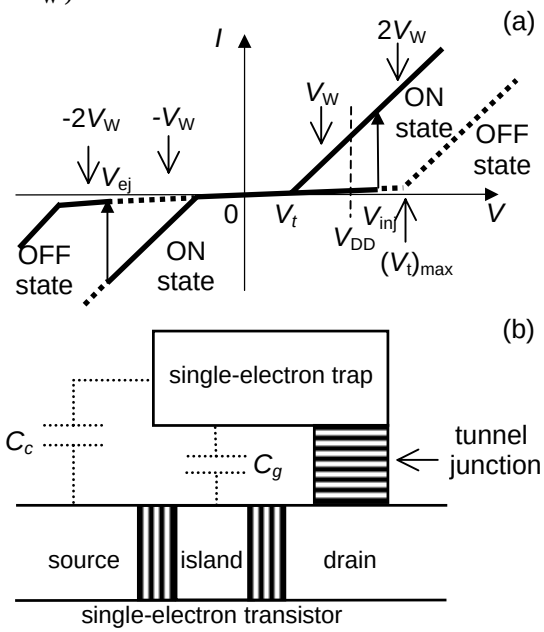

Fig. 2: Two-terminal latching switch: (a) The *I*-*V* curve (schematically) and (b) the single-electron implementation. In the OFF state of the switch, the single-electron transistor has a high Coulomb blockade threshold $(V_t)_{max} > V_{DD}$. If the source-drain voltage V exceeds a certain value $V_{inj} < (V_t)_{max}$, an additional electron is injected into the single-electron trap, and its electric field suppresses the Coulomb blockade threshold to a lower value $V_t < V_{DD}$, enabling the source-to-drain current flow. (The ON state of the latch.) The device may be turned OFF by applying a voltage below V_{ej} and thus ejecting the additional electron from the trap island.

Though low-temperature prototypes of such nanodevices have already been demonstrated (for a review, see Ref. 3), their fabrication yield is still low and will hardly ever reach 100%. This is why the most important requirement to hybrid circuit architectures is high defect tolerance. Recently we have shown that the CMOL approach allows to reach such tolerance, together with high performance, in digital terabit-scale memories [6] and FPGA-like digital logic circuits [7]. (For a recent review of these results, see Ref. 4.)

Even higher defect tolerance may be expected from neuromorphic architectures, due to their natural parallelism. We have suggested [8, 9] a family of such architectures, dubbed "CrossNets", that are naturally mapped on the CMOL circuit fabric. The goal of this presentation is to review the status of CrossNet development, with an emphasis on recent results.

2. CrossNets: FlossBar, InBar, etc.

Figure 3 shows the generic architecture of our Distributed Crossbar Networks ("CrossNets") on the example of its two main varieties, FlossBar and InBar. Relatively sparse neural cell bodies ("somas") are implemented in the CMOS subsystem. In the simplest firing rate model, each soma is just a differential amplifier with a nonlinear saturation ("activation") function $V_{out} = f(U_{in})$.

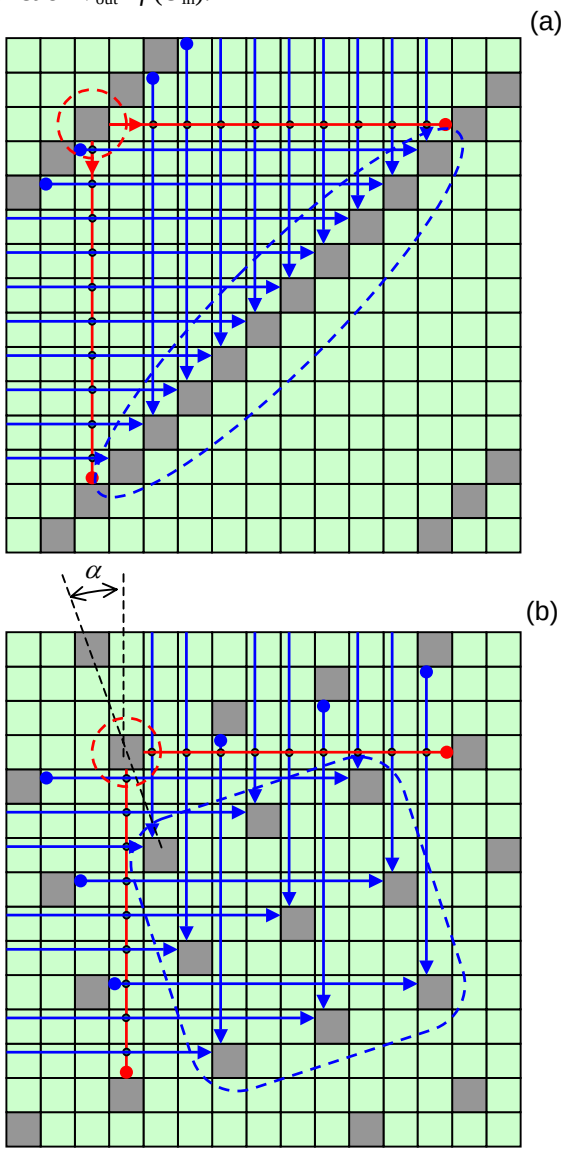

Fig. 3: Two main CrossBar species: (a) FlossBar and (b) InBar, in the simplest (feedforward, non-Hebbian, binary-weight) case. Only the nanowires and nanodevices coupling one cell (indicated with red dashed line) to its neighbors are shown. Axonic nanowires are shown in red, dendritic nanowires in blue. Each gray square show the interface pin area of a somatic cell. (The cells as such may be much larger, since they are implemented in the underlying CMOS subsystem.) In panel (a), the connectivity parameter M equals 9; in panel (b), $M = 16$.

Axons and dendrites are implemented as physically similar, straight segments of mutually perpendicular nanowires, while nanodevices (latching switches), formed at the nanowire crosspoints, play the role of elementary synapses. Each axonic voltage V_a, developed by the somatic amplifier, is transferred by an axonic nanowire to many (M) synapses, so that if a particular synaptic latch is in the ON state, a current proportional to V_a is flowing into the corresponding dendritic nanowire, contributing to the input signal of the post-synaptic cell. As a result, the dynamics of a feedforward CrossNet may be adequately described by the standard set of differential equations

$$(\tau M \frac{d}{dt} + 1) U_j = \sum_{k=1}^{M} w_{jk} V_k, \quad V_j = f(U_j), \qquad (1)$$

where the time constant τ is proportional to the specific capacitance of dendritic nanowires and the ON resistance of the latching switch. In the generic CrossNets (Fig. 3), any pair of cells is connected by two synapses leading to the opposite inputs of the somatic amplifier, so that the net synaptic weight w_{jk} may take any of three values which may be normalized to -1, 0, and +1. In other CrossNets versions the number of synapses is larger:

(i) In "Hebbian" CrossNets, each signal is passed, in the dual-rail format, through two nanowires, thus doubling the number of wires and making each pair of cells connected by two groups of 4 latching switches. The motivation for this modification is that the average net synaptic weight of the 4-switch group obeys the quasi-Hebbian law

$$\frac{d}{dt}\langle w \rangle = -4\Gamma_0 \sinh(\gamma S)\sinh(\gamma V_a)\sinh(\gamma V_d), \qquad (2)$$

where Γ_0 and γ are parameters of the latching switch, while S is a global, externally-controllable shift voltage which may be applied to all switches via a special gate.

(ii) In recurrent CrossNets, the number of nanowires and synapses per cell is doubled again, to carry feedback signals. (Generally, CrossNets are asymmetric: $w_{kj} \neq w_{jk}$.)

(iii) Each switch may be replaced for a block of $n \times n$ switches. This allows to get quasi-continuous synaptic weights with $L = 2n^2+1 >> 1$ values.

The most important parameter of CrossNet is their connectivity M. CMOL circuits, despite their quasi-2D structure, enable CrossNets with arbitrary values of M. Thus the CrossNets may span all the connectivity range, from the cellular automata (such as CNN) and global (e.g., Hopfield) networks.

Another important characteristic of a CrossNet is its cell location geometry. For example, FlossBars (Fig. 3a) are convenient for the implementation of multilayer perceptrons, while InBars (Fig. 3b) are more natural for the implementation of recurrent (in particular, quasi-Hopfield) networks.

3. CrossNet Training

CrossNet training as neural networks faces several hardware-imposed challenges:

(i) The synaptic weight contribution provided by the elementary latching switch is binary.

(ii) The only way to adjust any particular synapse is to apply certain voltage $V = V_a - V_d$ between the two corresponding nanowires, and this has to be done without disturbing semi-selected synapses.

(iii) Processes of turning single-electron latches ON and OFF are statistical rather than dynamical [3], so that the applied voltage V can only control probability rates of these random events.

We have shown [9] that all these challenges may be met using (at least) the following training methods.

3.1. Synaptic weight import

In this method, first, a homomorphic "precursor" artificial neural network with continuous synaptic weighs w_{jk} is trained using one of existing methods. Then the weights w_{jk} are transferred to the CrossNet, with some "clipping" (rounding) due to the finite number L of synaptic weight levels. This procedure is especially straightforward when w_{jk} may be simply calculated, e.g., for "clipped" (binary-weight) Hopfield networks [8].

3.2. Error backpropagation

For the price of a slight complication of the somatic cell, the quasi-Hebbian property expressed by Eq. (2) may be used to implement the standard backpropagation training algorithm in "Hebbian" CrossNets with multi-value synapses [9], without any special error propagation network. This mode seems especially useful for pattern classification applications.

3.3. Global reinforcement

We have obtained the first positive results for training of recurrent Hebbian InBars using a simple version of global reinforcement. For example, such complex task as the parity function has been successfully demonstrated (so far, for a small network with just three binary inputs). This method, however, still has to be extended to more complex tasks where an instant reward is not available after every output signal update.

4. Discussion

The developed methods of CrossNet training allow these networks to perform virtually any information processing functions that had been demonstrated with software-based artificial neural networks. The significance of this result is that the CMOL implementation may allow CrossNets to have extremely high performance. Indeed, estimates show [9] that CMOL CrossNets with connectivity even as high as 10^4 may reach an areal density of $\sim 3\times10^7$ cells/cm^2. This density corresponds to placing a system with the integration scale of the human brain on a 30×30 cm^2 silicon wafer. Moreover, the average signal delay per cell in such systems may be as low as 10 ns (at manageable power consumption), the number which should be compared with ~10 ms for the mammal cerebral cortex.[1] The neurobiology is still very far from teaching us how exactly such system should be organized. However, one may hope that the high speed may enable such (necessarily, hierarchical) system to self-develop, after a period of initial rudimentary supervised training, through broadband interaction with the environment [9].

The work has been supported by AFOSR, NSF, and MARCO (via FENA Center).

5. References

[1] C. Mead, *Analog VLSI and Neural Systems* (Addison-Wesley, Reading, MA), 1989.

[2] *International Technology Roadmap for Semiconductors. 2003 Edition, 2004 Update*, http//public.itrs.net/.

[3] K. K. Likharev, "Electronics Below 10 nm", *Nano and Giga Challenges in Microelectronics* (Elsevier, Amsterdam) pp. 27-68, 2003.

[4] K. K. Likharev and D. B. Strukov, "CMOL: Devices, Circuits, and Architectures", to be published in *Introduction to Molecular Electronics* (Springer, Berlin), 2005.

[5] S. Fölling, Ö. Türel, and K. K. Likharev, "Single-Electron Latching Switches as Nanoscale Synapses", *Proceedings of the 2001 IJCNN*, pp. 216-221, 2001.

[6] D. B. Strukov and K. K. Likharev, *Nanotechnology*, vol. 16, pp. 137-148, Jan. 2005.

[7] D. B. Strukov and K. K. Likharev, "CMOL FPGA", accepted for publication in *Nanotechnology*, 2005.

[8] Ö. Türel and K. K. Likharev, *Int. J. of Circ. Theor. Appl.*, vol. 31, pp. 37-53, 2003.

[9] Ö. Türel, J. H. Lee, X. Ma, and K. K. Likharev, *Int. J. of Circ. Theor. Appl.*, vol. 32, pp. 277-302, 2004.

[10] J. H. Lee and K. K. Likharev, "Defect Tolerance of CMOL CrossNet Classifiers", paper in preparation.

[1] Simultaneously, CrossNets may be very defect-tolerant, providing high fidelity at a fraction of bad nanodevices as high as 80% (!) in the quasi-Hopfield operation mode [9], and approximately 50% when working as pattern classifiers [10].