CrossNets: Neuromorphic Hybrid
CMOS/Nanoelectronic Networks

Konstantin K. Likharev
Department of Physics and Astronomy, Stony Brook University, Stony Brook, NY 11794-3800, USA

Hybrid CMOS/nanoelectronic circuits, combining CMOS chips with simple nanoelectronic crossbar add-ons, may extend the exponential Moore-Law progress of microelectronics beyond the 10-nm frontier. This paper reviews the development of neuromorphic networks (“CrossNets”) based on this prospective technology. In these networks, the neural cell bodies (“somas”) are implemented in the CMOS subsystem, crossbar nanowires are used as axons and dendrites, while two-terminal crosspoint devices are used as elementary synapses. Extensive analysis and simulations have shown that such networks may perform virtually all information processing tasks demonstrated with software-implemented neural networks, with much higher performance. Estimates show that CrossNets may eventually overcome bio-cortical circuits in density, at comparable connectivity, while operating 4 to 6 orders of magnitude faster, at manageable power dissipation.

Keywords: Nanoelectronics, Hybrid Circuits, Crossbar, Latching Switches, Memristive Devices, Neuromorphic Circuits, Neural Networks, Plasticity, Adaptation, Cognitive Tasks.

CONTENTS
1. Introduction: CMOL/Nanoelectronic Hybrids ................................ 322
2. Topologies .................................................................. 323
3. Performance Estimates ............................................. 325
4. Synaptic Weight Setting/Adaptation/Plasticity ...................... 326
5. Application Examples ................................................ 329
6. Prospects, Challenges, and Options .............................. 330
Acknowledgments ................................................................ 330
References and Notes ..................................................... 330

1. INTRODUCTION: CMOL/ NANOELECTRONIC HYBRIDS

Recent research has shown that the impending crisis of the exponential (“Moore’s-Law”) progress of microelectronics may be postponed for more than a decade by the transfer from purely CMOS technology to hybrid CMOS/nanodevice circuits.1-4 In such a circuit, a specially designed CMOS chip is complemented with a simple nanoelectronic add-on: a nanowire crossbar with simple, similar, two-terminal nanodevices at each crosspoint (Fig. 1). Two types on two-terminal nanodevices are being explored for use in the hybrids: “latching switches” (sometimes called “resistive switches” or “programmable diodes”), with resistive hysteresis shown schematically in Figure 2, and “memristive” devices whose parameters, including effective resistance, depend gradually of their operation history.4 Effective connection between the CMOS subsystem and the crossbar subsystem may be provided by an area-distributed interface; Figure 3 shows the so-called “CMOL” version of such interface,1-4,6 which allows the CMOS subsystem to address every nanowire, and hence every crosspoint device of the add-on crossbar. The basic idea behind such CMOS/nanoelectronic hybrids is that nanowire levels of the crossbar do not need alignment (“overlay”),5,8 and hence may be fabricated using advanced patterning methods, such as nanoimprint,9-11 EUV interference lithography,12-14 or block-copolymer lithography15 for whom the nanoscale-accurate overlay is not available. As a result, the crossbar half-pitch $F_{nano}$ may be much smaller than that ($F_{CMOS}$) of the CMOS subsystem.

Detailed simulations have shown that for the ratio $F_{CMOS}/F_{nano} \sim 10$, which may be anticipated by the end

---

*This author prefers to refer the currently popular term “memristors” to the original theoretical concept by Chua,* from which the experimentally demonstrated memristive devices differ rather substantially.

*Independently, a similar idea has been suggested by Cerofolini and Romano.*

*A prototype of a simplified version of the CMOL interface, called FPNI, has already been demonstrated experimentally.*
of this decade,23 the CMOS/nano hybrids with CMOS interface may provide a nearly two-orders-of-magnitude advantage over pure CMOS (with the same $F_{\text{CMOS}}$ and power per unit area, and comparable speed) for at least two digital applications: resistive memories24 and FPGA-like reconfigurable logic circuits.25–26 This advantage alone, equivalent crudely to 10 to 15 years of Moore’s Law extension, may be a sufficient motivation for the industrial introduction of the hybrids.

The goal of this paper is to review another area of possible applications of such hybrids in integrated circuits of another type: bio-inspired neuromorphic networks—“CrossNets,”2–5,27–35 whose estimated performance over CMOS circuits with similar functionality is even much higher, up to 6 orders of magnitude. The review starts with a description (in Section 2) of the basic idea and topology of CrossNets, followed by their performance estimates (Section 3). Section 4 reviews the developed methods for the import and/or runtime adaptation of their synaptic weights (which, for neural computation, play the role of programming at usual computing). Several examples of tasks which may be performed by CrossNets are described in Section 5, while the concluding Section 6 discusses prospects of their practical applications and major challenges to be met on the way towards their practical introduction.

2. TOPOLOGIES

Figure 4 shows the generic geometry of CrossNets. Neural cell bodies (somas) are implemented in the CMOS subsystem. For the simplest firing-rate models of neural networks,36–38 the soma may be just an analog amplifier with saturation, while in more bio-plausible, “spiking” models the somatic circuit receives, processes, and generates its own nerve pulses (“spikes”).

Output signal voltage $V_k(t)$ of a soma is applied, through an area-distributed interface (Fig. 3), to two “axonic” nanowires, in Figure 4 shown with red lines. Perpendicular, physically similar “dendritic” nanowires of the crossbar (blue lines) lead to inputs of other somas. If the two-terminal device at the crosspoint of $k$-th axon and $j$-th

Konstantin K. Likharev received the Candidate (Ph.D.) degree in Physics from the Lomonosov Moscow State University, Russia in 1969, and the habilitation degree of Doctor of Sciences from the Higher Attestation Committee of the U.S.S.R. in 1979. From 1969 to 1988 Dr. Likharev was a Staff Scientist of Moscow State University, and from 1989 to 1991 the Head of the Laboratory of Cryoelectronics of that university. In 1991 he assumed a Professorship at Stony Brook University (Distinguished Professor since 2002). During his research career, Dr. Likharev worked in the fields of nonlinear classical and dissipative quantum dynamics, and solid-state physics and electronics, notably including superconductor electronics and nanoelectronics. He is an author of more than 250 original publications, 75 review papers and book chapters, 2 monographs, and several patents. Dr. Likharev is a Fellow of the APS and IEEE. More detailed information is available online at http://rsfq1.physics.sunysb.edu/~likharev/personal/.
Fig. 3. CMOL interface between the CMOS subsystem and nanowire crossbar: (a) side view and (b) top view (schematically). The specific angle $\alpha = \arctan (r/r_{\text{ori}})$, where $r \approx F_{\text{CMOS}}/F_{\text{nano}}$ is an integer, of the tilt between the square mesh of interface pins and the crossbar, allows the CMOS system to address every nanowire of the crossbar, and hence every crosspoint device, individually.

Fig. 4. Topology of the simplest (feedforward, binary-weight) CrossNet. Red lines show “axonic”, and blue lines, physically similar “dendritic” nanowires. Gray squares show positions of two CMOS-based somatic circuits, with their interfaces with the nanowire crossbar shown in darker gray. Green circles denote latching switches forming elementary synapses. (For clarity, only those of them responsible for signal transfer between two somas $k$ and $j$ are shown.) Nanowire open-circuit terminations (shown without solid dots) do not allow somas to communicate in bypass of synapses.

Fig. 5. Signal forwarding in two CrossNet species: (a) FlossBar with connectivity $M = 10$ and (b) InBar with $M = 9$. For clarity, only the nanowires and nanodevices coupling one pre-synaptic cell (indicated with red dashed lines) to $M$ post-synaptic cells (blue dashed lines) are shown; actually all other cells are similarly coupled.

The voltage at the dendrite is in its ON state (Fig. 2), this voltage provides a substantial contribution to the current injection $I_j$ into $j$-th dendritic wire, so that in the linear approximation, and low input load, the total current is

$$I_j(t) = \sum_{k=1}^{M} \sum_{s_j=\pm1} s_j G_{jk} V_k(t)$$

where $G_{jk}$ is the crosspoint device conductance, and $s_j = \pm1$ is the input polarity. Equation (1) is exactly the key add–multiply operation pertinent to any artificial neural network (with the product $s_j G_{jk}$ playing the role of the synaptic weight $w_{jk}$) which consumes most computing resources at the network implementation in software run on general-purpose digital computers. In CrossNets, this is an analog operation which may be extremely fast—see the next section.
Perhaps the most important feature of CrossNets is that their connectivity $M$ (the upper limit in the sum in Eq. (1), i.e., the number of cells providing signal to any given one “directly” (via a synaptic contact), depends only on the distance between the somas and theoretically is unlimited, despite the quasi-2D geometry of these circuits. This is very important for modeling of (and eventually competing with) bio-cortical circuits whose average connectivity is close to $10^2$.39

Besides the intercell distance (and hence the connectivity), CrossNet properties depend on the cell distribution over the synaptic field. Figure 5 shows the feedforward versions of two CrossNet types most explored so far, the so-called FlossBar and InBar.28 The former network is more natural for the implementation of multilayered perceptrons (MLP, see Section 5.2), while the latter system may be preferable for recurrent network implementations (Section 5.1).

The generic topologies shown in Figures 4–5 may be readily extended to more advanced neuromorphic networks. For example, flexible combinations of FlossBar and InBar plaquettes are straightforward to engineer.55 Also, if the used crosspoint devices are bistable (Fig. 2), i.e. if a single device provides binary synaptic weight, synapses with multi-level weights may be organized from small arrays of such bistable switches (Fig. 6); two complementary square arrays, of $n \times n$ switches each, provide $L = n^2 + 1$ discrete weight levels, with $L = 33$ (i.e., $n = 4$) being sufficient for some key algorithms.32

3. PERFORMANCE ESTIMATES

The most important motivation for the architecture development, quantitative simulation (and eventually hardware implementation) of CMOL. CrossNets comes from estimates of their possible density, speed, and power. In CMOL topology, the total area occupied by synapses serving one cell is close to

$$A \approx 4 \frac{MLP^2}{nano}$$

At reasonable connectivity ($M > 10^2$) and weight level discreteness ($L \geq 33$), this area is more than sufficient for the layout of even the most complex (say, spiking) somatic circuitry, so that CrossNet density may be estimated from Eq. (2). For such realistic numbers as $L = 33, M = 3 \times 10^3$, and $F_{nano} = 5 \text{ nm}$, it yields $A \sim 10^{-7} \text{ cm}^2$, corresponding to approximately 100 M neural cells per cm$^2$. This is already close to mammal cerebral cortex neuron density (per cortex area$^{39}$). A substantial additional density increase may be obtained by the further crossbar scaling (some conceptual problems such as quantum-mechanical tunneling between the adjacent nanowires do not start until $\sim 2 \text{ nm}$), quasi-3D40 and genuine-3D41 integration, as well as using single “memristive” crosspoint devices, with continuously adjustable conductance, as synapses—see Section 6 below.

CrossNet speed estimates are even more impressive. Since some crosspoint devices may be switched faster than 100 ns (see Tables I–III in Ref. [6]), the main speed limitation comes from the RC time of dendritic nanowire recharging by ON currents of latching switches. With the calculated specific capacitance $C_0$ close to 0.2 $\text{fF/}$$\mu\text{m}_2$, a CrossNet with parameters cited above would have $C \approx C_0 A^{1/2} \sim 1 \text{ pF}$. Concerning the effective resistance $R \approx R_{ON}/2 \text{ ML}$, crosspoint devices with $R_{ON}$ values within a broad range have been demonstrated, so that the most important limitation comes from the necessity to keep the circuit power density $P \approx V^2/8R_{nano}R_{ON}$ (where $V$ is the axonic voltage scale) at a manageable level. With $V = 1 \text{ V}$, the parameters cited above yield $P \approx 1 \text{ W/cm}^2$ at $R_{ON} \approx 0.5 \times 10^{12} \Omega$ ($R \approx 2.5 \times 10^9 \Omega$). For this power level, which does not require dedicated cooling, the intercell delay scale $RC$ is of the order of 2.5 $\mu$s. If $P$ is increased to 100 W/cm$^2$ (typical for the high-performance microprocessors), the latency decreases to $\sim 25$ ns. These numbers are, respectively, approximately 3 and 5 orders of magnitude lower than the average intercell latency in the biological cortex.38, 39

Of course, such comparison of CrossNet with biocortical circuitry would be completely fair only if their functionality had been close. So far, theoretical neuroscience is still very far from telling us how this goal may be achieved, and gives recipes for performing only relatively simple cognitive tasks (some of which, nevertheless, already have important practical applications). The next section describes some options for implementation such recipes in CrossNets.
4. SYNAPTIC WEIGHT SETTING/ADAPTATION/PLASTICITY

In neurocomputing, the role of programming is played by setting synaptic weights to their desired values—either before the calculation or in runtime. (Weight adaptation in the latter mode is frequently referred to as network “plasticity.”) What follows is the list of procedures which have already been developed and demonstrated (on adequate computer models of CrossNets) for this purpose.

4.1. Weight Import

If the set of synaptic weights for a particular task has been determined using an external digital computer (a “precursor”), these weights may be imported into a CrossNet. This task is not so trivial. Indeed, setting a certain set \( w_{ij} \) of synaptic weights in a CrossNet with bistable crosspoint devices requires switching each device into its proper state, either ON and OFF. For that, voltages applied to two nanowires leading to each device have to be manipulated in a way which ensures the desirable switching event in the selected device while not perturbing the states of other (“semi-selected”) devices connected to each of the wires. Moreover, the import procedure should be parallelized as much as possible to ensure practicable weight import times.

Nevertheless, such import procedures, with a number of time steps scaling as \( M \) (rather than the total number of crosspoint devices in the network!) have been developed both for InBars and FlossBars, both with binary\(^{29}\) and multi-level\(^{30,32}\) synaptic weights. There is a feeling that these solutions may be extended to virtually any future CrossNet topology.

4.2. In-Situ Adaptation: Firing-Rate Models

If the task of synaptic weight calculation is too large for a digital precursor, it has to be performed within the CrossNet itself. Of several adaptation algorithms, the Hebb rule and its variations\(^{36-38}\) are believed to be most important. Figure 7 shows the method (based on the well-known “statistical multiplication” approach) of providing the Hebb-type plasticity in firing-rate CrossNets.\(^{30}\)

In this method, each synapse consists of four arrays of \( n \times n \) elementary latching switches, fed by bipolar (dual-rail) voltages, so that in the signal propagation mode, the synaptic weight may take any of the \( L = 4n^2 + 1 \) quantized, equidistant values within the zero-centered range \([−w_{\text{max}}, +w_{\text{max}}]\). In the weight adaptation mode, voltages \( V_1(t) \) and \( V_2(t) \) are developed by comparators \( C_{1,2} \) with binary outputs. The comparators are fed by:

(i) the analog signals \( x_{1,2} \) (supplied by somatic cells), with magnitudes limited to a certain range \([0, x_{\text{max}}]\), and the signs changed in time to perform 4-stage time division multiplexing (see the table inset in Fig. 7); and

(ii) reference signals \( \text{REF}_{1,2} \) from two independent pseudo-random signal generators, with the uniform probability within the same range \([0, x_{\text{max}}]\).

In this arrangement, the probability of each comparator to apply the voltage of proper polarity to each output wire is proportional to the input analog signal. A straightforward analysis\(^{30}\) shows that, as a result, the average synaptic weight change during a time-division multiplexing cycle is

\[
\Delta w = \eta x_1 x_2 \times \begin{cases} (w_{\text{max}} - w), & \text{for } x_1 x_2 > 0 \\ (w_{\text{max}} + w), & \text{for } x_1 x_2 < 0 \end{cases}
\]

where \( \eta \) is a constant depending on the bistable device parameters and the comparator output voltage. This is just the standard \( x_1 x_2 \) form of the Hebb rule, but with additional saturation (which is unavoidable at hardware implementation of synaptic weights).

An important feature of this adaptation method is that all necessary circuitry (including the comparator and pseudo-random number generator) serving one axonic or dendritic line may be shared by all \( M \) synapses of that line. As a result, the overhead of CMOS hardware necessary for their implementation does not affect density of CrossNets with biologically-plausible values of connectivity \( M \).

4.3. In-Situ Adaptation: STDP

In spiking neuromorphic networks, which explicitly model neural pulses in biological systems, the most popular way of the Hebb rule implementation is the so-called spike-time-dependent plasticity, STDP.\(^{42,43}\) The STDP requires

\[
\]
to increase the probability of OFF $\rightarrow$ ON switching if the pre-synaptic and post-synaptic spikes are within a certain time window (the latter spike follows the former one), and suppress the probability in the opposite case.

Figure 8 shows an example of how this can be achieved in spiking CrossNets. Each spike consists of two voltage pulses: a longer pulse A and a shorter but higher pulse B, with the amplitude approaching (but somewhat below) the switching threshold $V_s$. Applied to a composite synapse (Fig. 6), the former pulse creates a current pulse (negative) polarity. Now the net voltage $V$ is increased, so that when the short pulse B is applied to the axonic nanowire, $V$ exceeds the threshold, creating a finite probability of ON switching for those crosspoint devices of the composite synapse (Fig. 8), for relatively high connectivity $M$. Small black points: raw results of 20,000 numerical experiments with random spike timing; red squares: values of $V_{\text{max}}$ averaged within $2.5 \tau_0$—wide time bins. Other parameters: $f_\text{fs} = 0.01$, $\tau_0/\tau_0 = \tau_1/\tau_0 = 10$, $(\tau_1 - \tau_1)/\tau_0 = 0.1$, $V_{\text{max}}/V_s = 2$.

Unfortunately, numerical modeling shows that this scheme does not promise good scaling of STDP plasticity with growing connectivity $M$ (Fig. 9). The reason is...
Table I. Approximate bias conditions for crosspoint synapses based on standard NAND flash memory cells.

<table>
<thead>
<tr>
<th>Electrode potentials (versus the p-well)</th>
<th>$V_{A_k}$</th>
<th>$V_{d_k}$</th>
<th>$V_{G_j}$</th>
<th>$V_{A_k}$</th>
<th>$V_{d_j}$</th>
<th>$V_{G_j}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spike propagation with STDP...</td>
<td>0 V</td>
<td>float</td>
<td>0 V</td>
<td>+5 V</td>
<td>float</td>
<td>+5 V</td>
</tr>
<tr>
<td>Block reset</td>
<td>0 V</td>
<td>0 V</td>
<td>−15 V</td>
<td>0 V</td>
<td>0 V</td>
<td>−15 V</td>
</tr>
<tr>
<td>Particular weight import ($G_{a} \leftarrow G_{max}$)</td>
<td>0 V</td>
<td>float</td>
<td>+15 V</td>
<td>+7 V</td>
<td>float</td>
<td>0 V</td>
</tr>
</tbody>
</table>

that at the bio-plausible average rate $f$ of spikes generated by each cell may be as high as $\sim 0.1/\tau$, where $\tau$ is the spike duration, so that the product $M/\tau$ may be much higher than 1, meaning that at the input of each soma, many spikes may overlap. As a result, the STDP response becomes noisy, and its average deviates from the desired antisymmetric function of the spike delay.

Ways toward better scaling still have to explored; here, let me only mention that these complications may be naturally avoided in “Flash CrossNets”-model circuits using flash memory cells working in the analog mode—see Figure 10 and Table I. (Earlier suggestions of using flash technology in neuromorphic networks were based on more complex cells.)

Figure 11 shows a typical result of theoretical analysis of such flash synapses, with the simple somatic feedback circuit shown in Figure 10(b), using both Monte Carlo modeling and an approximate quasi-analytical, single-spike approximation (exact in the limit $M/\tau \rightarrow 0$). The results indicate very good scaling of the STDP response even at $M/\tau \gg 1$, limited only by the biologically-plausible condition $f \tau < 1$, due to a natural separation of the adjustment feedback signals $V_g$ from feedforward dendritic signals $V_d$.

Of course, the flash memory technology is essentially a twist of CMOS, so that it requires patterning with accurate layer alignment, and cannot be scaled down as much as nanowire crossbars. However, in Flash CrossNets, one flash cell may provide the synaptic weight accuracy comparable to that of a multi-latching-switch array (Figs. 6, 7), at a comparable network density. (Suggestions to use continuously-adjusted memristive crosspoint devices for providing analog synaptic weights would probably require much lower device-to-device variability than the one demonstrated experimentally.)

5. APPLICATION EXAMPLES

The results presented in this section have been obtained by CrossNet simulation using their realistic models. They give some idea about possible performance of such networks.

5.1. Hopfield Networks: Pattern Recognition

Possibly the simplest type of an artificial neural net is the recurrent, firing-rate network with symmetric synaptic weights, $w_{jk} \equiv w_{kj}$. (Such networks had been explored by several researchers before they were made famous by Hopfield.) Properly trained, the Hopfield network may work as an associative memory, using a part of a pre-written patterns to restore (“recognize”) the whole pattern.

Since the capacity of such memory is very weakly affected by synaptic weight discreteness, a CrossNet with just one latching switch per synapse may operate very well in this mode; its main difference from the generic Hopfield net is the quasi-local (rather than global) connectivity $M$, limiting its capacity to $\sim 0.45 M$ at 99% restoration fidelity.

Figure 12 shows an example of such an operation; the final image is completely error free. However, the most remarkable feature of the pattern restoration is its speed ($\sim 5 RC$), taking into account that in realistic CrossNets the $RC$ time constant may be below 1 $\mu$s. (See Section 3 above.)

Fig. 12. Dynamics of recognition of one of three trained black-and-white images by an InBar-type CrossNet with 256 × 256 neural cells and connectivity parameter $M = 64$. The initial image (left panel) was obtained from the trained image (identical to the one shown in the right panel) by flipping as many as 40% of pixels at random. $RC$ is the effective time constant of intercell interaction. Reprinted with permission from [53], Ö. Türel and K. K. Likharev, Possible nanoelectronic implementation of neuromorphic networks, Proc. IJCNN’03 (2003), pp. 365–370. © 2003, IEEE.
5.2. Multilayer Perceptrons: Pattern Classification

A much more important, but also more demanding function of neuromorphic networks is the pattern classification which may be achieved, for example, in layered perceptrons after their supervised training by error backpropagation—see, e.g., Refs. [36–38]. Two major concerns about using CrossNets in this mode have been: (i) the necessary accuracy of synaptic weights, and (ii) defect tolerance.

Figure 13 shows typical results of study,32 which used a very common benchmark—the MNIST database of typewritten characters.54 It shows that, for example, \( L = 33 \) synaptic levels (available from two \( 4 \times 4 \) composite synapses shown in Fig. 6) are sufficient for getting virtually the same fidelity (\( \sim 98\% \)) as for exact (continuous) synaptic weights, and that a very substantial number of stuck-at-closed defects cause only a slow fidelity degradation.

These results pertain to weights imported from a precursor network; this training method is quite sufficient, for example, for a known face recognition in a large crowd, because it may be based on using multiple copies of the desired pattern, with a TV-raster-type search (Fig. 14). Estimates have shown55 that such a CrossNet chip, with area below 1 cm\(^2\), may identify a face on a 8-Mpixel image in approximately 100 \( \mu \)s, the number to be compared with \( \sim 10^3 \) s for the same algorithm run on a general-purpose microprocessor.

5.3. Global Reinforcement and TD Learning

Some cognitive tasks require unsupervised learning, in particular, global reinforcement with either instant or delayed reward.36–38 A study of this mode of CrossNet operation has shown that these networks are quite suitable for the most popular global-reinforcement algorithms, such as \( A_\text{ri} \).56 Moreover, they may use similar algorithms (which have been called \( A_1 \) and \( A_{1/\sigma} \)) based on synaptic rather than somatic randomness, which are more natural for nanodevice implementation of synapses. Figure 15 shows that these new algorithms provide just a slightly lower learning speed than \( A_\text{ri} \) for the cart-and-pole balancing task—a popular benchmark for global reinforcement with delayed reward.

6. PROSPECTS, CHALLENGES, AND OPTIONS

Though studies of possible CrossNet applications are in the very beginning, it looks like that these networks may be used for performing virtually any cognitive task which had been demonstrated using software-implemented neural nets, at very high speed (with manageable power...

References and Notes


Received: 29 December 2010. Accepted: 14 January 2011.