CMOL

Jung Hoon Lee, Xialong Ma, Dmitri B. Strukov and Konstantin K. Likharev *
Stony Brook University, NY 11794-3800, U.S.A.
*Email: klikharev@notes.cc.sunysb.edu

ABSTRACT

This is a brief review of the recent work on architectures for the prospective hybrid semiconductor/nanodevice ("CMOL") integrated circuits. The basic idea of such circuits is to combine the advantages of the currently dominating CMOS technology (including its flexibility and high fabrication yield) with those of molecular-scale nanodevices with nanometer-scale footprint, at acceptable fabrication costs. Preliminary estimates show that the density of active devices in CMOL circuits may be as high as 10^{12} cm^{-2} and that they may provide an unparalleled information processing performance, up to 10^{20} operations per cm^2 per second, at manageable power consumption. The most straightforward application of CMOL circuits is terabit-scale memories, in which powerful bad-bit-exclusion and error-correction techniques may be used to boost the defect tolerance. Our preliminary results for reconfigurable, FPGA-like digital-logic CMOL circuits also look very encouraging. Finally, CMOL technology seems to be uniquely suitable for the implementation of the "CrossNet" family of neuromorphic networks for advanced information processing including, at least, pattern recognition and classification, and quite possibly much more intelligent tasks.

1. INTRODUCTION

The recent results [1, 2] indicate that the current VLSI paradigm, based on a combination of lithographic patterning, CMOS circuits, and Boolean logic, can hardly be extended into a-few-nm region. The main reason is that at gate length below 10 nm, the sensitivity of parameters (most importantly, the gate voltage threshold) of silicon MOSFETs to inevitable fabrication spreads grows exponentially. As a result, the gate length should be controlled with a few-angstrom accuracy, far beyond even the long-term projections of the semiconductor industry [3]. Even if such accuracy could be technically implemented using sophisticated patterning technologies, this would send the fabrication facilities costs (growing exponentially even now) skyrocketing, and lead to the end of the Moore's Law during the next decade.

The main alternative nanodevice concept, single-electronics [2, 4], offers some potential advantages over CMOS, including the scalability to atomic dimensions and a broader choice of possible materials. Unfortunately, for room-temperature operation the minimum features of these devices (single-electron islands) should be below $\sim$1 nm [4]. Since the relative accuracy of their definition has to be between 10 and 20%, the absolute fabrication accuracy should be of the order of 0.1 nm, again far too small for the current and realistically envisioned lithographic techniques.

This is why there is a rapidly growing consensus that the impending crisis of the microelectronics progress may only be resolved by a radical paradigm shift from the lithography-based fabrication to the "bottom-up" approach. In the latter approach, the smallest active devices should be formed in a special way ensuring their fundamental reproducibility. The most straightforward example of such a device is a specially designed and chemically synthesized molecule of a few hundred atoms, including functional parts (e.g., acceptor groups working as single-electron islands, and short fragments of non-conducting groups as tunnel junctions), and the groups enabling chemically-directed self-assembly of the molecule on prefabricated electrodes.[1] The recent experimental demonstration of molecular single-electron transistors, based on such approach, by several groups [5–9] has shown that it is viable. We believe that the further development of this approach may lead to the practical introduction, perhaps within the next 10 to 20 years, of the first integrated circuits with active nanodevices.

Unfortunately, integrated circuits consisting of molecular devices alone are hardly practicable, because of limited device functionality. For example, voltage gain of a 1-nm-scale transistor, based on any known physical effect (e.g., the field effect, quantum interference, or single-electron charging), can hardly exceed one [2], i.e. the level necessary for sustaining the operation of virtually any active analog or digital circuit.[2] This is why we believe that the only plausible way toward high-performance nanoelectronic circuits is to integrate molecular devices and the connecting nanowires with CMOS circuits whose (relatively bulky) field-effect transistors would provide the necessary additional functionality, in particular high voltage gain.

Recently, several specific proposals of such circuits were published. (A detailed review of this, and some other previous work on molecular electronics circuitry may be found in Ref. 11.) The goal of this paper is to review the recent work in one promising direction toward hybrid semiconductor-molecular electronics, the

[1] Plus, most likely, some additional groups ensuring sufficient rigidity and stability of the molecule at room temperature.

[2] The very recent suggestion [10] to replace transistors with the Goto pairs of two-terminal latching switches in crossbar circuits runs into several problems, most importantly the relation between the retention time and switching speed. In order to be useful for most electronics applications, the latches should be switched very fast (in a few picoseconds in order to compete with advanced MOSFETs), but retain their internal state for the time necessary to complete the calculation (ideally, for a few years, though several hours may be acceptable in some cases). This means that the change of the applied voltage by the factor of two (the difference between the fully selected and semi-selected crosspoints of a crossbar) should change the switching rate by at least 16 orders of magnitude. However, even the most favorable physical process that we are aware of (the quantum-mechanical tunneling through high-quality dielectric layers like the thermally-grown SiO_2) may only produce, at these conditions, the rate changes below 10 orders of magnitude, even if uncomfortably high voltages of the order of 10 V are used.

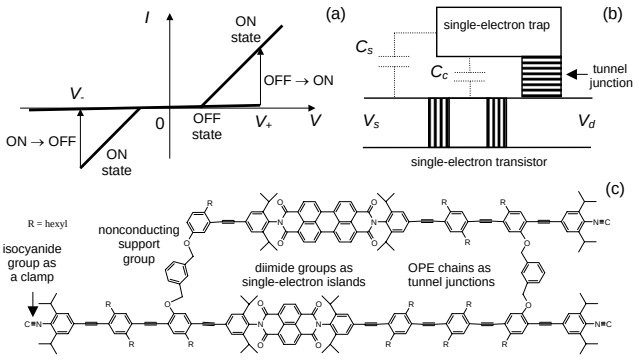

Figure 1: Two-terminal latching switch: (a) $I - V$ curve (schematically), (b) single-electron device schematics [13], and (c) a possible molecular implementation of the device (courtesy A. Mayr).

so-called CMOL approach.

2. DEVICES

The first critical issue in the development of semiconductor/ nanodevice hybrids is making a proper choice in the trade-off between nanodevice simplicity and functionality. Our group's preference is the binary "latching switch", i.e. a two-terminal, bistable device with $I - V$ curves of the type shown in Fig. 1a.[3] Such switch may be readily implemented, for example, as a combination of two single-electron devices: a "transistor" and a "trap" (Fig. 1b).[4] If the applied drain-to-source voltage $V = V_d - V_s$ is low, the trap island in equilibrium has no extra electrons ($n = 0$), and its net electric charge $Q = -ne$ is zero. As a result, the transistor is in the virtually closed (OFF) state, and the source and drain are essentially disconnected. If V is increased beyond a certain threshold value V_+, its electrostatic effect on the trap island potential (via capacitance C_s) leads to tunneling of an additional electron into the trap island: $n \to 1$. This change of trap charge affects, through the coupling capacitance C_c, the potential of the transistor island, and suppresses the Coulomb blockade threshold to a value well below V_+. As a result, the transistor, whose tunnel barriers should be thinner than that of the trap, is turned into ON state in which the device connects the source and drain with a finite resistance R_0. (Thus, the trap island plays the role similar to that of the floating gate in the usual nonvolatile semiconductor memories.) If the applied voltage stays above V_+, this state is sustained indefinitely; however, if V remains low for a long time, the trapped electron eventually leaks out, and the transistor is closed, disconnecting the electrodes. This ON $\to$ OFF switching may be forced to happen much faster by making the applied voltage V sufficiently negative, $V \approx V_-$.

Figure 1c shows a possible molecular implementation of the device shown in Fig. 1b. Here two different diimide acceptor groups play the role of single-electron islands, while short oligo-ethynylene-phenylene (OPE) chains are used as tunnel barriers. The chains are terminated by isocyanide-group "clamps" ("alligator clips") that should enable self-assembly of the molecule across a gap between two metallic electrodes. It is important that several (many) such molecules connected in parallel would work similarly to one device, besides an increased current scale.

The potentially enormous density of nanodevices can hardly be used without individual contacts to each of them. This is why the fabrication of wires with nanometer-scale cross-section is another central problem of molecular microelectronics. The currently available photolithography methods, and even their rationally envisioned extensions, will hardly be able to provide such resolution. Several alternative techniques, like the direct e-beam writing and scanning-probe manipulation can provide a nm-scale resolution, but their throughput is forbiddingly low for VLSI fabrication. Self-growing nanometer-scale-wide structures like carbon nanotubes or semiconductor nanowires can hardly be used to solve the wiring problem, mostly because these structures (in contrast with the specially synthesized molecules that have been discussed above) do not have means for reliable placement on the lower integrated circuit layers with the necessary (a-few-nm) accuracy. Fortunately, there are several new patterning methods, notably nanoimprint [14] and interference lithography [15], which may provide very high resolution (in future, down to a few nanometers) compared to the standard photolithography.

3. CIRCUITS

These novel patterning technologies cannot be used, however, for the fabrication of arbitrary integrated circuits, in particular because they lack adequate layer alignment accuracy. This means that the nanowire layers should not require precise alignment with each other and with the CMOS subsystem. While the former requirement may be readily satisfied by using the "crossbar" nanowire structure (i.e., two layers of similar wires perpendicular to those of the other layer), the solution of the latter problem (CMOS-to-nanowire interface) is much harder. In fact, the interface should enable the CMOS subsystem, with a relatively crude device pitch $2\beta F_{\text{CMOS}}$ (where $\beta \sim 1$ is the ratio of the CMOS cell size to the wiring period), to address each wire separated from the next neighbors by a much smaller distance (nanowiring half-pitch) F_{nano}.

Several solutions to this problem, which had been suggested earlier, seem either unrealistic, or inefficient, or both. In particular, the interface based on statistical formation of semiconductor-nanowire field-effect transistors gated by CMOS wires [16, 17] can only provide a limited (address-decoding-type) connectivity. In addition, the resistivity of semiconductor nanowires would be too high for high-performance hybrid circuits. Even more importantly, the technology of ordering chemically synthesized semiconductor nanowires into highly ordered parallel arrays has not been developed, and the authors of this paper are not aware of any promising idea that may allow such assembly.

A more interesting approach [11] is to cut off the ends of nanowires of a parallel-wire array, along a line that forms a small angle $\alpha = \arctan(F_{\text{nano}}/F_{\text{CMOS}})$ with the wire direction. As a result of the cut, the ends of adjacent nanowires stick out by distances (along the wire direction) differing by $2F_{\text{CMOS}}$, and may be contacted individually by the similarly cut CMOS wires. Unfortunately, the latter (CMOS) cut has to be precisely aligned with the former (nanowire) one, and it is not clear how exactly such a feat might be accomplished using available patterning techniques.

Figure 2 shows our approach to the interface problem. (We call such circuits "CMOL", standing for CMOS/nanowire/MOLecular-scale-device hybrids.) The conceptual difference between the CMOL approach (based on earlier work on the so-called "InBar" networks [18, 19]), and the earlier suggestions discussed above [11] is that in CMOL the CMOS-to-nanowire interface is provided by pins dis-

[3] Multi-terminal devices would be immeasurably more complex for implementation, e,g, for the chemically-directed molecular self-assembly.

[4] Low-temperature prototypes of this device, with a slightly different design of the trap, have been implemented and successfully tested to provide electron trapping times beyond 12 hours [12].

tributed all over the circuit area.[5] In the generic CMOL circuit (Fig. 2), pins of each type (contacting the bottom and top nanowire levels) are located on a square lattice of period $2\beta F_{\rm CMOS}$. Relative to these arrays, the nanowire crossbar is turned by a (typically, small) angle α which satisfies two conditions (Fig. 2b):

$$\sin\alpha = F_{\rm nano}/\beta F_{\rm CMOS}, \quad (1)$$

$$\cos\alpha = rF_{\rm nano}/\beta F_{\rm CMOS}, \quad (2)$$

where r is a (typically, large) integer. Such tilt ensures that a shift by one nanowire (e.g., from the second wire from the left to the third in Fig. 2c) corresponds to the shift from one interface pin to the next (in the next row of similar pins), while a shift by r nanowires leads to the next pin in the same row. This trick enables individual addressing of each nanowire even at $F_{\rm nano} \ll \beta F_{\rm CMOS}$. For example, the selection of CMOS cells 1 and 2 (Fig. 2c) enables contacts to the nanowires leading to the left of the two nanodevices shown on that panel. Now, if we keep selecting cell 1, and instead of cell 2 select cell 2' (using the next CMOS wiring row), we contact the nanowires going to the right nanodevice instead.

It is also clear that if all the nanowires and nanodevices are similar to each other (the assumption that will be accepted in all the following discussion), a shift of the nanowire/nanodevice subsystem by one nanowiring pitch with respect to the CMOS base does not affect the circuit properties. Moreover, a straightforward analysis of Fig. 2c shows that at an optimal shape of the interface pins, even a complete lack of alignment of these two subsystems leads to a circuit yield loss about 75%. Such loss may be acceptable, taking into account that the cost of the nanodevice fabrication (e.g., molecular self-assembly) may be rather low, especially in the context of an unparalleled density of active devices in CMOL circuits. In fact, the only evident physical limitation of the density is the quantum-mechanical tunneling between parallel nanowires. Simple estimates show that the tunneling current becomes substantial at the distance between the wires $F_{\rm nano} \approx 1.5$ nm. Even by accepting a more conservative value of 3 nm, we get the device density $n = 1/(2F_{\rm nano})^2$ above $10^{12}{\rm cm}^{-2}$, i.e. at least three orders of magnitude higher than any purely CMOS circuit ever tested.

4. CMOL MEMORIES

The similarity of all nanodevices, which seems necessary for the simplicity of CMOL circuit fabrication, imposes substantial restrictions on architectures, and hence possible applications of the circuits. An even more essential restriction comes from the anticipated finite yield of nanodevice fabrication, which will hardly ever approach 100%. As a result, all practical CMOL architectures should be substantially defect-tolerant.

This tolerance may be most simply implemented in embedded memories and stand-alone memory chips, with their simple matrix structure. In such memories, each nanodevice (for example the single-electron latching switch - see Fig. 1) would play the role of a single-bit memory cell, while the CMOS subsystem may be used for coding, decoding, line driving, sensing, and input/output functions.[6]

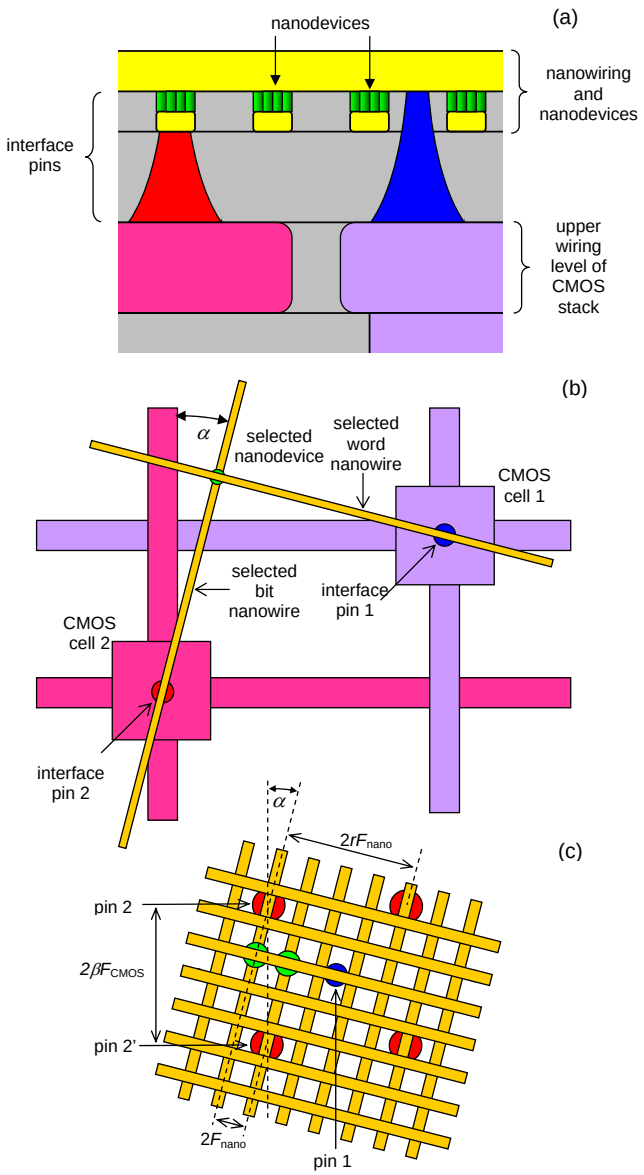

Figure 2: The generic CMOL circuit: (a) a schematic side view; (b) a schematic top view showing the idea of addressing a particular nanodevice via a pair of CMOS cells and interface pins, and (c) a zoom-in top view on the circuit near several adjacent interface pins. On panel (b), only the activated CMOS lines and nanowires are shown, while panel (c) shows only two devices. (In reality, similar nanodevices are formed at all nanowire crosspoints.) Also disguised on panel (c) are CMOS cells and wiring.

We have carried out [22] a detailed analysis of one such memory, using one of two known techniques for increasing its defect tolerance: (i) the memory matrix reconfiguration (the replacement of several rows and columns, with the largest number of bad memory

[5] Such sharp-pointed pins may be fabricated similarly to the tips used in field-emission arrays - see, e.g., Ref. 20.

[6] It may seem that a large problem in such memories is the necessity for the latching switches to combine a sufficient retention time and write/erase speed (see Footnote 2). However, in memories the speed requirements may be substantially relaxed: a-few-microsecond write/erase time may be acceptable for some, and a-few-nanosecond time, for most applications. Moreover, the periodic memory refresh (similar to that used in the present-day DRAM) may allow to use cells with retention time as low as a few seconds. Hence, the switching speed ratio (at the doubling of applied voltage) should be from about 5 to 9 orders of magnitude. The former requirement may be easy to satisfy, while the latter challenge may possibly be met using single-electron trap barriers with an appropriate structure [21].

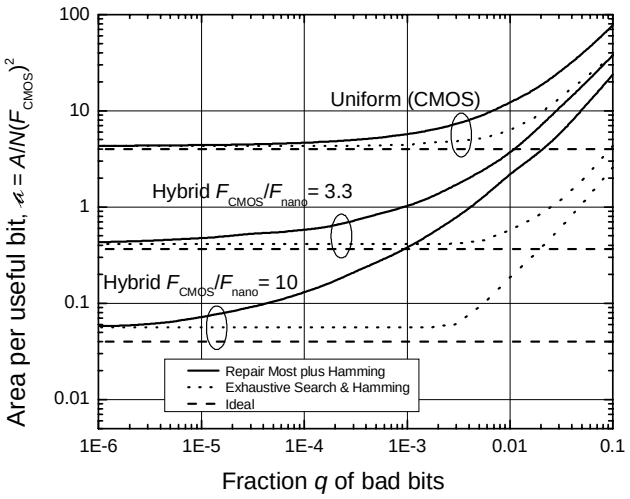

Figure 3: The area per useful bit after the block size optimization, as a function of single bit yield, for hybrid and purely semiconductor memories.

cells, for spare lines), and error correction (based on the Hamming codes). The analyzed memory was a matrix of L memory blocks, each block in turn being a rectangular array of $(n + a) \times (m + b)$ memory cells. Here a and b are the numbers of spare rows and columns, respectively, while $n \times m$ is the final block size after the reconfiguration. A p-bit word addressed at each particular time step is distributed over p blocks. Each block is a global crossbar matrix. At each elementary operation, the block decoders address two vertical and two horizontal lines implemented in the CMOS layers of the circuit, thus selecting a pair of CMOS cells (Fig. 2b). Each cell has a simple "relay" structure (using either one or two pass transistors [22]) and connects one of the CMOS-level wires leading to the cell to the interface pin, and hence to the corresponding nanowire. As has been explained in Sec. 3 above, this allows the four cell address decoders of each block to reach each memory cell, even if the cell density is much higher than $1/(F_{\rm CMOS})^2$.

We have considered several combinations of the Hamming-code error correction with two reconfiguration techniques[7]:

(i) a simple "Repair Most" reconfiguration algorithm, in which a worst rows of the array (with the largest number of bad bits) are excluded first, and b worst columns of the remaining matrix next; and

(ii) the best possible, but exponentially complex "Exhaustive Search" reconfiguration.

The simulation results show that the array reconfiguration ("repair") improves the memory yield rather dramatically, while the difference between the two repair methods is not too large, especially if the number of redundant lines is not too high - below, or of the order of the final memory size. (The difference is somewhat larger if the array reconfiguration is used together with the error correction.)

We have evaluated the additional memory area necessary to get a certain fixed yield, as a function of the memory parameters, in particular the block size (at fixed total memory size). The area is contributed by spare lines necessary for the array configuration, additional parity bits necessary for the Hamming-code error correction, and CMOS components including the decoders, drivers, sense amplifiers and a relatively small CMOS-based memory storing the reconfiguration results.

Figure 3 shows the total chip area per useful bit, optimized over the block size, as a function of the nanodevice yield, for two values of the $F_{\rm CMOS}/F_{\rm nano}$ ratio and two defect tolerance boost techniques. (Results for purely CMOS memories are also shown for comparison.) The results show that the density CMOL memories may be very impressive. For example, the normalized cell area $a \equiv A/N(F_{\rm CMOS})^2$ = 0.4 (Fig. 3) at $F_{\rm CMOS}$ = 32 nm means that a memory chip of a reasonable size (2×2 cm^2) can store about 1 terabit of data - crudely, one hundred Encyclopedia Britannica's.

However, the defect tolerance results are not too encouraging. For example, in a realistic case $F_{\rm CMOS}/F_{\rm nano} = 10$, the hybrid memories can overcome a perfect CMOS memory only if the fraction of bad bits is below $\sim 15\%$, even using the Exhaustive Search algorithm of bad bits exclusion, which may require an impracticably long time. For the simple and fast Repair Most algorithm, the bad bit fraction should be reduced to $\sim 2\%$. If one wants to obtain an order-of-magnitude density advantage from the transfer to hybrid memories (such a goal seems natural for the introduction of a novel technology), the numbers given above should be reduced to approximately 2% and 0.1%, respectively. These results show that the global-crossbar approach to CMOL memory architectures is not too promising. Presently we are working on a more distributed architecture which promises much higher defect tolerance, at comparable useful bit density.

5. CMOL FPGA: RECONFIGURABLE LOGIC CIRCUITS

In the usual digital logic circuits the location of a defective gate from outside is hardly possible. On the other hand, the incorporation of additional logic gates (e.g., providing von Neumann's majority multiplexing) for error detection and correction becomes very inefficient for a fairly low fraction q of defective devices. For example, even the recently improved von Neumann's fault exclusion scheme requires a 10-fold redundancy for q as low as $\sim 10^{-5}$ and a 100-fold redundancy for $q \approx 3 \times 10^{-3}$ [23].

This is why the most significant previously published proposals for the implementation of logic circuits using CMOL-like hybrid structures had been based on reconfigurable regular structures like the field-programmable gate arrays (FPGA). Before our recent work, two FPGA varieties had been analyzed, one based on look-up tables (LUT) and another one using programmable-logic arrays (PLA). Unfortunately, all these approaches run into substantial problems - see, e.g., Ref. 24 for their critical review.

Recently, we suggested [25] an alternative approach to Boolean logic circuits based on the CMOL concept, which reminds the so-called cell-based FPGA [26]. In this approach (Fig. 4a, b), an elementary CMOS cell includes two pass transistors and one inverter, and is connected to the nanowire/ molecular subsystem via two pins.[8] During the configuration process the inverters are turned off, and the pass transistors may be used for setting the binary state of each molecular device, just like described above for CMOL memory. Each pin of a CMOS cell can be connected through a

[7] Assuming so far only one type of defects: the absence of nanodevices at certain crosspoints, formally equivalent to the "stuck-on-open" faults.

[8] For convenience of signal input and output, the nanowire crossbar is turned by additional 45° in comparison with the generic CMOL (Fig. 2), so that Eqs. (1), (2) now take the form $\sin\alpha = (r - 1)F_{\rm nano}/\beta F_{\rm CMOS}$, $\cos\alpha = rF_{\rm nano}/\beta F_{\rm CMOS}$. Also note the breaks in each nanowire in the middle of its contacts with the interface pins.

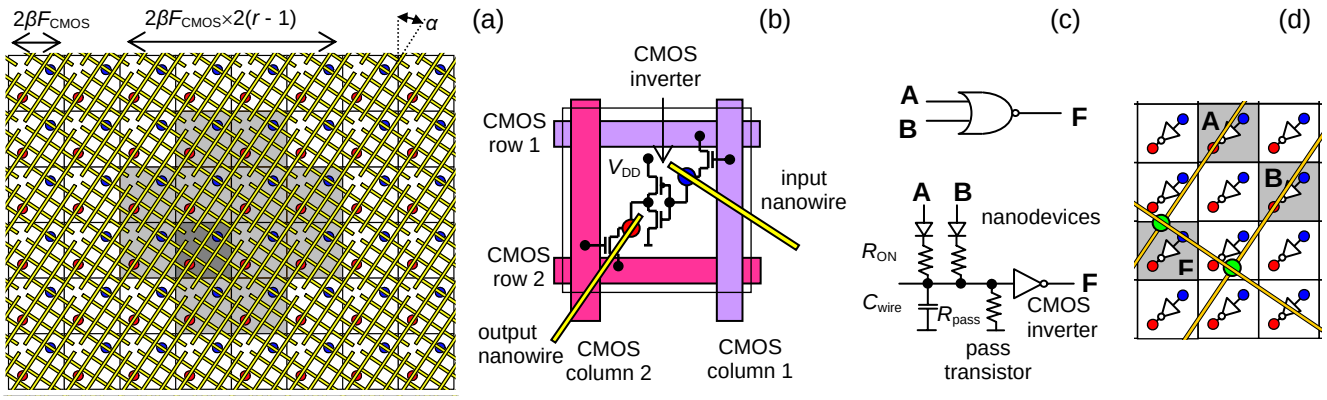

Figure 4: CMOL FPGA: (a) the general structure of the circuit and (b) a single CMOS cell, and (c) NOR gate implementation. In panel (a), the cells painted light-gray may be connected to the input pin of a specific cell (shown dark-gray). For the sake of clarity, panel (b) shows only two nanowires (that contact the given cell), while panel (c) shows only the three nanowires used inside the NOR gate.

nanowire-nanodevice-nanowire link to each of $M \equiv 2r^2 - 2r - 1$ other cells within a square-shaped "connectivity domain" around the pin (painted light-gray in Fig. 4a). Figure 4c shows how such fabric may be configured for the implementation of a fan-in-two NOR gate. This is already sufficient to implement any logic function, though gates with larger fan-in and fan-out are clearly possible.

Note that during the circuit operation the switching latches should not change their state, i.e. work either as diodes, if they are in their ON state, or open circuits with some (parasitic) high resistance, if they are turned OFF (Fig. 1a). This is why the switching speed to retention time requirement (see Footnote 2) is relaxed even more than in CMOL memories: while the retention time should be long (at least a few hours, or better yet, a few years), the programming time as long as a few seconds may be acceptable, since the programming of the whole circuit requires just $\sim M$ sequential steps.

Generally, there may be many different algorithms to reconfigure the CMOL FPGA structure around known defects. We have developed a CMOL FPGA configuration procedure which is carried out in two stages. First, the desired circuit is mapped on the apparently perfect (defect-free) CMOL fabric.[9] At the second stage, the circuit is reconfigured around defective components using a simple algorithm, linear in M [25]. Our Monte Carlo simulation (again, so far only for the "no-assembly"-type defects) has shown that even this simple configuration procedure may ensure very high defect tolerance. For example, the reconfiguration of a 32-bit Kogge-Stone adder, mapped on the CMOL fabric with realistic values of parameters $r = 12$ and $r' = 10$, may make a system with as many as 50% of missing nanodevices fully functional. Under the requirement of a 99% circuit yield (sufficient for a 90% yield of properly organized VLSI chips), the defect tolerance of this circuit is about 22%, while that of another key circuit, a fully-connected 64-bit crossbar switch, is about 25%. These impressive results may be explained by the fact that each CMOS cell is served by $M \gg 1$ nanodevices used mostly for reconfiguration.

It is especially important that CMOL FPGA circuits combine such high defect tolerance with high density and performance, at acceptable power consumption. Indeed, approximate estimates have

[9] We have found it very beneficial, for boosting defect tolerance, to confine the cell connections to a smaller square shaped domain of $M' \equiv 2r'^2 - 2r' - 1$ cells, with r' slightly below the maximum connectivity radius r.

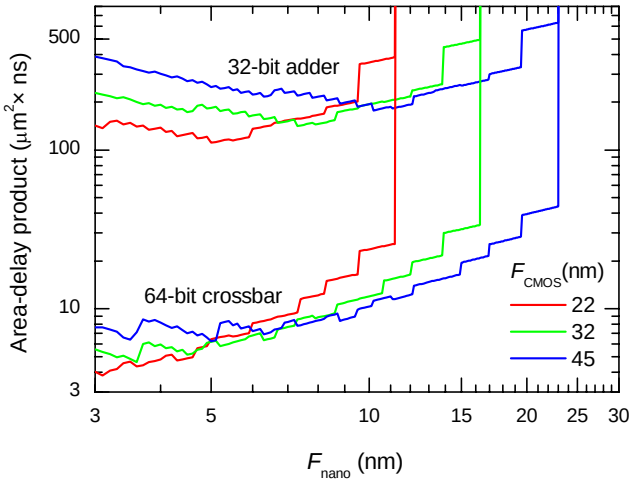

Figure 5: CMOL FPGA area-delay product optimization results as functions of nanowire half-pitch $A\tau$ of the two CMOL FPGA circuits for three ITRS long-term CMOS technology nodes. The (formal) jump of the $A\tau$ product to infinity at some $(F_{\text{nano}})_{\text{max}}$ reflects the fact that circuit mapping on the CMOL fabric may only be implemented for F_{nano} below this value. The finite sharp jumps of the curves are due to the discrete changes of angle α.

shown [25] that for the power of 200 W/cm^2 (planned by the ITRS for the long-term CMOS technology nodes [3]), an optimization of the power supply voltage V_{DD} may bring the logic delay of the 32-bit Kogge-Stone adder down to just 1.9 ns, at the total area of 110 μm^2, i.e. provide an area-delay product of 150 ns-μm^2, for realistic values $F_{\text{CMOS}} = 32$ nm and $F_{\text{nano}} = 8$ nm (Fig. 5). This result should be compared with the estimated 70,000 ns-μm^2 (with 1.7 ns delay and 39,000 μm^2 area) for a fully CMOS FPGA implementation of the same circuit (with the same F_{CMOS} = 32 nm).

In the sample circuits explored so far, each CMOS cell is using just not more than two latching switches for actual operation. Recently we have got a preliminary indication that an increase of this number to four or five may increase the circuit performance twofold, hopefully with a negligible sacrifice of the defect toler-

ance. A quantitative study of this opportunity is one of our immediate goals.

However, a final conclusion on the potential value of CMOL FPGA circuits may be only made after their defect tolerance and performance have been evaluated for a substantial number of various functional units and other circuits necessary for digital signal processing and/or general-purpose computing, using generally accepted computing benchmarks.

6. CMOL CROSSNETS: NEUROMORPHIC NETWORKS

The requirement of high defect tolerance gives a strong incentive to consider CMOL implementation of alternative information processing architectures, in particular analog or mixed-signal neuromorphic networks (see, e.g., Ref. 27), since such networks are by their structure deeply parallel and hence inherently defect-tolerant. An additional motivation for using neuromorphic networks comes from the following comparison between the performance of the biological neural systems and present-day digital-logic computers in one of the basic advanced information processing tasks: image recognition (more strictly speaking, classification [27]). A mammal's brain recognizes a complex visual image, with high fidelity, in approximately 100 milliseconds. Since the elementary process of neural cell-to-cell communication in the brain takes approximately 10 milliseconds, it means that the recognition is completed in just a few "clock ticks". In contrast, the fastest modern microprocessors performing digital number crunching at a clock frequency of a few GHz and running the best commercially available code, would require many hours (i.e., of the order of 10^{13} clock periods) for an inferior classification of a similar image. The contrast is very striking indeed, and serves as a motivation for the whole field of artificial "neural networks".

Presently, these networks are mostly just a concept for writing software codes that are implemented on usual digital computers. Unfortunately, the high expectations typical for the neural network's "heroic period" (from the late 1980s to the early 1990s) have not fully materialized, in particular because the computer resources limit the number of neural cells to a few hundreds, insufficient for performing really advanced, intelligent information processing tasks. The advent of hybrid CMOL circuits may change the situation.

Recently, our group suggested [18, 19] a new family of neuromorphic network architectures, Distributed Crosspoint Networks ("CrossNets" for short) that map uniquely on the CMOL topology. Each such network consists of the following components:

(i) Neural cell bodies ("somas") that are relatively sparse and hence may be implemented in the CMOS subsystem. Most of our results so far have been received within the simplest Firing Rate approach [27], in which somas operate just as differential amplifiers with a nonlinear saturation ("activation") function, fed by the incoming (dendritic) nanowires, which apply their output signal to outcoming (axonic) wires.

(ii) "Axons" and "dendrites" that are implemented as mutually perpendicular nanowires of the CMOL crossbar.

(iii) "Synapses" that control coupling between the axons and dendrites (and hence between neural cells) based on the molecular latching switches (see Fig. 1 and its discussion).

CrossNet species differ by the number and direction of intercell couplings (Fig. 6), and by the location of somatic cells on the axon/dendrite/synapse field (Fig. 7). The cell distribution pattern determines the character of cell coupling. For example, the "FlossBar" (Fig. 7a) has a layered structure typical for the so-called

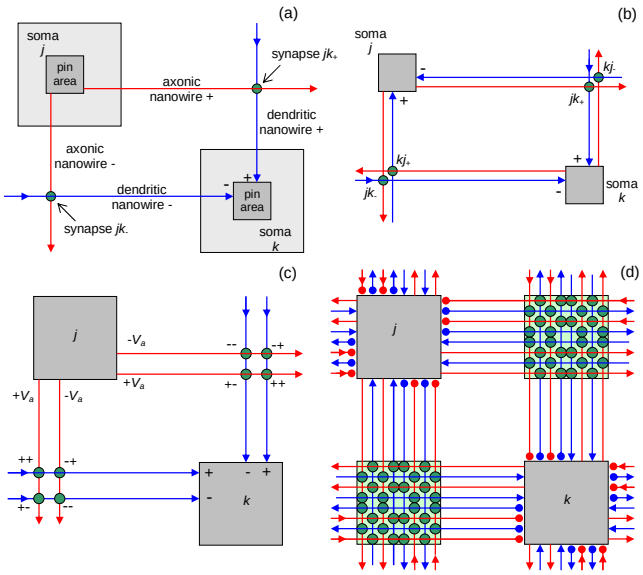

Figure 6: Schemes of cell connections in CrossNets: (a) simple (non-Hebbian) feedforward network, (b) simple recurrent network, (c) Hebbian feedforward CrossNet and (d) Hebbian recurrent CrossNet [28]. Red lines show "axonic", and blue lines "dendritic" nanowires. Dark-gray squares are interfaces between nanowires and CMOS-based cell bodies (somas), while light-gray squares in panel (a) show the somatic cells as a whole. (For the sake of clarity, the latter areas are not shown in the following panels and figures.) Signs show the somatic amplifier input polarities. Green circles denote nanodevices (latching switches) forming elementary synapses. For clarity the panels (a)-(c) show only the synapses and nanowires connecting one couple of cells (j and k). In contrast, panel (d) shows not only those synapses, but also all other functioning synapses located in the same "synaptic plaquettes" (painted light-green) and the corresponding nanowires, even if they connect other cells. (In CMOL circuits, molecular latching switches are also located at all axon/axon and dendrite/dendrite crosspoints; however, they do not affect the network dynamics, resulting only in approximately 50% increase of power dissipation.) The solid dots on panel (d) show open-circuit terminations of synaptic and axonic nanowires, that do not allow direct connections of the somas, in bypass of synapses.

multilayered perceptrons [27], while the "InBar" (in which somas sit on a square lattice inclined by a small angle relatively the axonic/dendritic lattice, Fig. 7b) implements a non-layered "interleaved" network. Also important is the average distance M between the somas, which determines connectivity of the networks, i.e. the average number of other cells coupled directly to a given soma. The most remarkable property of CMOL CrossNets is that the connectivity of these (quasi-)2D structures may be very large. This property is very important for advanced information processing, and distinguishes CrossNets favorably from the so-called cellular automata with small (next-neighbor) connectivity which severely limits their functionality.

In contrast to the usual computers, neuromorphic networks do not need an external software code, but need to be "trained" to perform certain tasks. For that, the synaptic connections between the cells should be set to certain values. The neural network science has developed several effective training methods [27]. The applica-

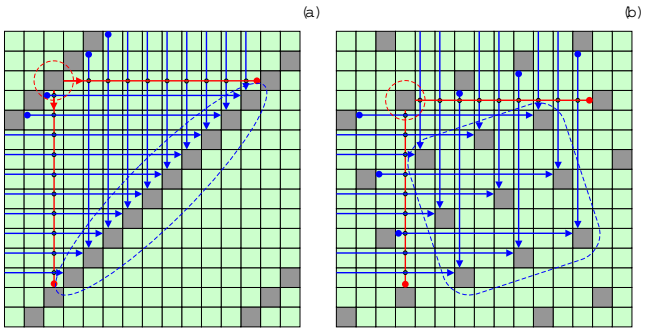

Figure 7: Two particular CrossNet species: (a) FlossBar and (b) InBar. For clarity, the figures show only the axons, dendrites, and synapses providing connections between one soma (indicated by the dashed red circle) and its recipients (inside the dashed blue lines), for the simple (non-Hebbian) feedforward network.

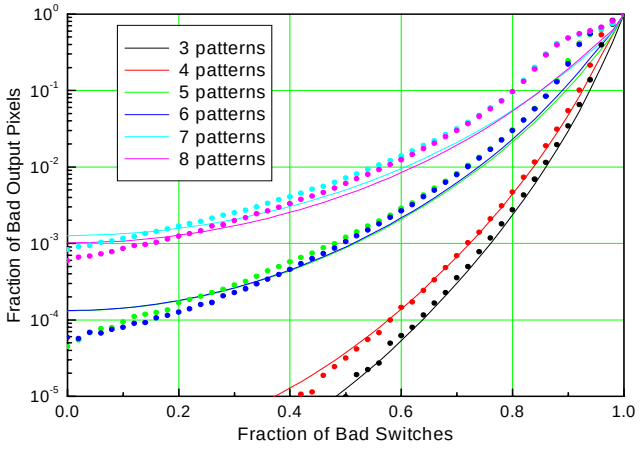

Figure 8: Defect tolerance of a recurrent InBar with connectivity parameter $M = 25$, operating in the quasi-Hopfield mode. Lines show the results of an approximate analytical theory, while dots those of a numerical experiment.

tion of these methods to CMOL CrossNets faces several hardware-imposed challenges:

(i) CrossNets use continuous (analog) signals, but the synaptic weights are discrete (binary, if only one latching switch per synapse is used).

(ii) The only way to reach for any particular synapse in order to turn it on or off is through the voltage V applied to the device through the two corresponding nanowires. Since each of these wires is also connected to many other switches, special caution is necessary to avoid undesirable "disturb" effects.

(iii) Processes of turning single-electron latches on and off are statistical rather than dynamical [4], so that the applied voltage V can only control probability rates Γ of these random events.

In our recent work [28] we have proved that, despite these limitations, CrossNets can be taught, by at least two different methods, to perform virtually all the major functions demonstrated earlier with usual neural networks, including the corrupted pattern restoration in the recurrent quasi-Hopfield mode and pattern classification in the feedforward multilayered perceptron mode [28, 29].[10] Moreover, at least in the former mode the CrossNets can be spectacularly resilient. For example, operating at network capacity just a half of its maximum, a quasi-Hopfield CrossNet may provide a 99% result fidelity with as many as 85% (!) of bad molecular devices - see Fig. 8. This defect tolerance is much higher than that of CMOL FPGA circuits (Sec. 5).

The fact that CrossNets may perform the tasks that had been demonstrated with artificial neural networks earlier, may not seem very impressive until the possible performance of this hardware is quantified. Estimates [18, 19, 28] show that for realistic parameters as have been used in Sec. 4 above ($F_{\rm nano}$ = 4 nm, V = 0.25 Volt), and a very respectable connectivity parameter $M \sim 10^3$, the areal density of CrossNets may be at least as high as that of the cerebral cortex (above 10^7 cells per cm^2), while the average cell-to-cell communication delay τ_0 may be as low as $\sim$ 10 ns (i.e., about six orders of magnitude lower than in the brain), at power dissipation below 100 W/cm^2.[11] This implies, for example, that a 1-cm^2 CMOL CrossNet chip would be able to recognize a specific in a high-resolution image (e.g., a certain face in a crowd) faster than in 100 microseconds [29]. We believe that such applications alone may form not just a narrow market niche, but a substantial market for the hybrid CMOS/molecular electronics.

More speculatively, there is a hope that CrossNets will be able to perform even more complex intelligent tasks if trained using more general methods such as global reinforcement [27, 30]. (We have already obtained some preliminary positive results for such training [28].) If these hopes materialize, there will be a chance that such pre-training of a properly organized, hierarchical CrossNet-based system, may enable functionality comparable with that of a newborn child's brain, in some sense replacing the DNA-based genetic inheritance. It seems possible that a connection of such pre-trained system to a proper informational environment via a high-speed communication network may trigger a self-development process that may be several orders of magnitude faster than that of the biological cerebral cortex. The reader is invited to imagine possible consequences of such self-development liberated from the dead weight artifacts of the biological evolution.

7. CONCLUSIONS

There is a chance for the development, perhaps within the next 10 to 20 years, of hybrid "CMOL" integrated circuits which would allow an extension of the Moore's Law to the few-nm range. Preliminary estimates show that such circuits could be used for several important applications, notably including terabit-scale memories, reconfigurable digital circuits with multi-teraflops-scale performance, and mixed-signal neuromorphic networks that may, for the first time, compete with biological neural systems in areal density, far exceeding them in speed, at acceptable power dissipation. We believe that these prospects justify large-scale multi-disciplinary research and development efforts in the fields of synthesis of func-

[10] In order to operate as perceptron-type classifiers, CrossNets require multi-latch synapses. This increase can be achieved by using small (e.g., 4 $\times$4) square fragments of CrossNet arrays for each synapse [28]. This increase is taken into account in the density estimates given below.

[11] The reason for such a large difference with power estimates for Boolean logic circuits (Sec. 5) is that in neuromorphic networks we can afford to increase the open molecular latch resistance to $\sim 10^9$ ohms, and thus increase the logic delay from $\sim$100 ps to $\sim$10 ns, still providing an extremely high integrated circuit performance ($\sim 10^{12}/10^{-8} \approx 10^{20}$ of a-few-bit operations per cm^2 per second) due to the natural parallelism of the neuromorphic network operation.

tional molecular devices, their chemically-directed self-assembly, nanowire patterning, and CMOL circuit architectures.

Acknowledgment

Useful discussions of the issues considered in this paper with P. Adams, J. Barhen, V. Beiu, W. Chen, E. Cimpoiasu, S. Das, J. Ellenbogen, X. Liu, J. Lukens, A. Mayr, V. Protopopescu, M. Reed, M. Stan, and Ö. Türel are gratefully acknowledged. Figure 1c is a courtesy by A. Mayr. The work on this topic at Stony Brook was supported in part by AFOSR, NSF, and MARCO via FENA Center.

8. REFERENCES

[1] D. J. Frank, *et al.*, "Device scaling limits of Si MOSFETs and their application dependencies," *Proc. IEEE*, vol. 89, no. 3, pp. 259–288, 2001.

[2] K. K. Likharev, "Electronics below 10 nm," in *Nano and Giga Challenges in Microelectronics*. Amsterdam: Elsevier, 2003, pp. 27–68.

[3] *International Technology Roadmap for Semiconductors. 2003 Edition, 2004 Update*, available online at http://public.itrs.net/.

[4] K. K. Likharev, "Single-electron devices and their applications," *Proc. IEEE*, vol. 87, no. 4, pp. 606–632, 1999.

[5] H. Park, *et al.*, "Nanomechanical oscillations in a single-C-60 transistor," *Nature*, vol. 407, no. 6800, pp. 57–60, 2000.

[6] S. P. Gubin, *et al.*, "Molecular clusters as building blocks for nanoelectronics: The first demonstration of a cluster single-electron tunnelling transistor at room temperature," *Nanotechnology*, vol. 13, no. 2, pp. 185–194, 2002.

[7] N. B. Zhitenev, H. Meng, and Z. Bao, "Conductance of small molecular junctions," *Phys. Rev. Lett.*, vol. 88, no. 22, p. 226801, 2002.

[8] J. Park, *et al.*, "Coulomb blockade and the Kondo effect in single-atom transistors," *Nature*, vol. 417, no. 6890, pp. 722–725, 2002.

[9] S. Kubatkin, *et al.*, "Single-electron transistor of a single organic molecule with access to several redox states," *Nature*, vol. 425, no. 6959, pp. 698–701, 2003.

[10] P. J. Kuekes, D. R. Stewart, and R. S. Williams, "The crossbar latch: Logic value storage, restoration, and inversion in crossbar circuits," *J. Appl. Phys.*, vol. 97, no. 3, p. 034301, 2005.

[11] M. R. Stan, *et al.*, "Molecular electronics: From devices and interconnect to circuits and architecture," *Proc. IEEE*, vol. 91, no. 11, pp. 1940–1957, 2003.

[12] L. Ji, *et al.*, "Fabrication and characterization of single-electron transistors and traps," *J. Vac. Sci. Technol. B*, vol. 12, no. 6, pp. 3619–3622, 1994.

[13] S. Fölling, Ö. Türel, and K. K. Likharev, "Single-electron latching switches as nanoscale synapses," in *International Joint Conference on Neural Networks*. Mount Royal, NY: Int. Neural Network Soc., 2001, pp. 216–221.

[14] S. Zankovych, *et al.*, "Nanoimprint lithography: Challenges and prospects," *Nanotechnology*, vol. 12, no. 2, pp. 91–95, 2001.

[15] S. R. J. Brueck, "There are no fundamental limits to optical lithography," in *International Trends in Applied Optics*. Bellingham, WA: SPIE Press, 2002, pp. 85–109.

[16] P. J. Kuekes and R. S. Williams, "Demultiplexer for a molecular wire crossbar network (MWCN DEMUX)," U.S. Patent 6 256 767, July 3, 2001.

[17] A. DeHon, P. Lincoln, and J. E. Savage, "Stochastic assembly of sublithographic nanoscale interfaces," *IEEE Trans. Nanotechnol.*, vol. 2, no. 3, pp. 165–174, 2003.

[18] Ö. Türel and K. Likharev, "CrossNets: Possible neuromorphic networks based on nanoscale components," *Int. J. Circ. Theory App.*, vol. 31, no. 1, pp. 37–53, 2003.

[19] K. Likharev, *et al.*, "CrossNets - High-performance neuromorphic architectures for CMOL circuits," *Ann. NY Acad. Sci.*, vol. 1006, pp. 146–163, 2003.

[20] K. L. Jensen, "Field emitter arrays for plasma and microwave source applications," *Phys. Plasmas*, vol. 6, no. 5, pp. 2241–2253, 1999.

[21] K. K. Likharev, "Riding the crest of a new wave in memory," *IEEE Circuits Devices Mag.*, vol. 16, no. 4, pp. 16–21, July 2000.

[22] D. B. Strukov and K. K. Likharev, "Prospects for terabit-scale nanoelectronic memories," *Nanotechnology*, vol. 16, pp. 137–148, 2005.

[23] S. Roy and V. Beiu, "Multiplexing schemes for cost-effective fault-tolerance," *IEEE Trans. Nanotechnol.*, 2005, accepted for publication.

[24] K. K. Likharev and D. B. Strukov, "CMOL: Devices, circuits, and architectures," in *Introducing Molecular Electronics*, G. Cuniberti, G. Fagas, and K. Richter, Eds. Berlin: Springer, 2005, to be published as Chapter 16.

[25] D. B. Strukov and K. K. Likharev, "CMOL FPGA: A cell-based, reconfigurable architecture for hybrid digital circuits using two-terminal nanodevices," *Nanotechnology*, submitted for publication, preprint available at http://rsfq1.physics.sunysb.edu/figures//nano/FPGA05.pdf.

[26] J. M. Rabaey, A. P. Chandrakasan, and B. Nikolic, *Digital integrated circuits : A design perspective*, 2nd ed. Upper Saddle River, NJ: Pearson Education, 2003.

[27] J. Hertz, R. G. Palmer, and A. S. Krogh, *Introduction to the theory of neural computation*. Redwood City, CA: Addison-Wesley Pub. Co., 1991.

[28] Ö. Türel, *et al.*, "Neuromorphic architectures for nanoelectronic circuits," *Int. J. Circ. Theory App.*, vol. 32, no. 5, pp. 277–302, 2004.

[29] J. H. Lee and K. K. Likharev, "CMOL CrossNets as pattern classifiers," in *8th International Work-Conference on Artificial Neural Networks*. Barcelona, Spain: Int. Neural Network Soc., 2005, preprint available at http://rsfq1.physics.sunysb.edu/figures//nano/IWANN05.pdf.

[30] R. S. Sutton and A. G. Barto, *Reinforcement learning : An introduction*. Cambridge, MA: MIT Press, 1998.