A Case for CMOS/nano Co-design

Matthew M. Ziegler and Mircea R. Stan
University of Virginia, ECE Department, Charlottesville, VA 22904
{ziegler, mircea}@virginia.edu

Abstract—
The challenge of extending Moore's Law past the physical and economic barriers of present semiconductor technologies calls for novel nanoelectronic solutions. Circuits composed of mixed silicon semiconductors and nanoelectronics can provide a means for gradually switching technology paradigms. We suggest a design methodology to accompany this concept. Furthermore, we explore design tradeoffs for a nanoscale crossbar technology that supports CMOS/nano co-design.

I. Introduction

While traditional silicon electronics should continue industrial dominance for at least the next decade, novel nanoelectronic solutions will be needed to surmount the physical and economic barriers of current semiconductor technologies and continue along the exponential projections of Moore's Law. Although most new nanoelectronic solutions are still in their infancy, they present the potential for unprecedented levels of device density, low power computing, and possibly higher operating speed. Despite this high potential, it will be very difficult for any new technology to compete head-to-head with silicon's large-scale fabrication infrastructure, proven design methodologies, and economic predictability.

For brevity, in this paper we refer to conventional silicon electronics, including future variations to silicon MOSFET-based electronics, as "CMOS". Likewise, we use the term "nano" to refer to novel nanoscale electronics.

A feasible scenario is that the exponential returns of silicon scaling will flatten about at the same time as nanoelectronics will mature towards high levels of integration. However, the prerequisites for any nanotechnology taking silicon's top spot in the electronics industry include a large industrial backing, mass fabrication abilities, adequate design methodologies and tools, possibly new architectures, plus many other obstacles. While these obstacles can be surmounted in the long term, an abrupt technology change is not likely to happen soon.

An alternative approach to an abrupt technology change is the integration of silicon with nanoelectronics, i.e. mixed CMOS/nano integrated circuits. This route would allow a smooth transition and permit leveraging the beneficial aspects of both technologies. A smooth transition can be achieved by first integrating a small amount of nano on a predominantly CMOS chip. In successive generations, the amount of nano can be increased as the amount of CMOS is decreased. Increasing the nano-to-CMOS ratio over time can provide a means to ease into a new technology paradigm. Furthermore, the possibility of mixed CMOS/nano circuits permits using the best aspects of both technologies simultaneously, while the undesired aspects of a technology can be compensated by the partner technology. The mixed CMOS/nano concept also will encounter a number of obstacles; however, the hope is that many of the undesirable obstacles encountered in an abrupt technology switch can be avoided.

0-7803-7607-2/02/$17.00

II. Mixed CMOS/nano Circuits

CMOS-based electronics will also require many breakthroughs in order for the semiconductor industry's ITRS roadmap to [1] hold true. Thus, even if CMOS a decade from now looks very different from today, we can estimate with a reasonable level of accuracy its status at that time. On the other hand, the future of nanoelectronics is much more difficult to project.

Currently there are a number of technologies that have high potential for nanoelectronics [2]. One key differentiating feature of any nanotechnology is whether the underlying fabrication approach is a "top-down" subtractive method or "bottom-up" self-assembly. Many nanotechnologies using top-down approaches, such as silicon and heterojunction resonant tunneling diodes (RTDs), show good performance. However, the physical dimensions of these devices will be limited by the resolution of the top-down process, e.g., lithography or nanoimprinting. On the other hand, the size limits of bottom-up self-assembly could be much smaller, since assembly is controlled on the atomic or molecular scale.

In this paper we focus on nanotechnologies that employ bottom-up methods, such as chemical self-assembly. One such technology, molecular nanoelectronics comprises devices and/or wires consisting of single or just a few molecules and represents nearly the ultimate limit in scaling. Bottom-up approaches typically cannot replicate the complex structures that top-down fabrication methods, such as lithography, can achieve. Thus, molecular circuits presently are restricted to regular or periodic structures that can be produced via self-assembly. The inherent tradeoff involves sacrificing arbitrary design complexity for a higher-density, regularly structured and potentially low-cost approach. A number of nanotechnologies based on regular structures have shown preliminary success [3], [4], [5].

While much of the analysis in this paper is applicable to several nanotechnology approaches based on regularly structured circuits, we use the crossbar technology proposed by Hewlett-Packard and UCLA for demonstrating ideas at a lower-level of abstraction [3], [4], [6], [7], [8]. This crossbar technology is composed of arrays of crossed nanoscale wires with bistable nanoscale switches sandwiched between the intersections of the wires. The upper-left portion of Fig. 1 shows a simplified diagram of such a crossbar. Molecules are present at each junction, forming a two-terminal device that can be electrically configured to behave as a low resistance diode or a high resistance diode. These molecules, such as rotaxanes or cantenanes, create a programmable computing fabric that can be used for memories, logic arrays, etc. Harvard has demonstrated a similar circuit paradigm consisting of crossed nanowire p-n junctions [9] as well as logic gates from crossed nanowire field-effect transistors (cNW-FETs) [5]. Conceptually, magnetic RAM (MRAM) also consists of a similar array of crossed wires with bistable junctions [10].

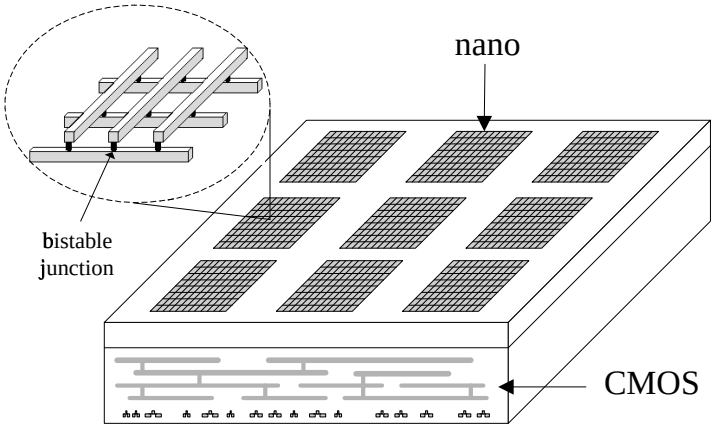


Fig. 1. We consider a design paradigm involving nanoelectronics on a CMOS IC.

Molecular crossbar technologies present an opportunity for high computational densities; however, they do suffer from some significant drawbacks. Most of the suggested bistable devices are two-terminal devices having the ability to switch between a low resistance state and a high resistance state. Rectifying (diode-like) behavior is also generally present in these devices, depending on the technology. Being restricted to a diode-resistor logic style results in an inability to achieve signal gain. The lack of gain will restrict the array size and require interfacing to technology capable of achieving gain for extended computation. Another drawback of diode-resistor logic is the inability to implement an inverter. Thus, without a complete logic family, both a signal and its complemented signal are required for full logic capabilities. The lack of inversion also complicates sequential storage elements. In addition, while the crossbar paradigm has a feasible route towards digital computation, developing analog crossbar circuits should prove more difficult. In general, realizing analog nano circuits will be complicated by signal noise and the inherent defect rates associated with bottom-up assembly. Therefore it seems that such crossbar technologies will require some sort of partner technology to adequately perform computation. Previously proposed nano-architectures suggest mixing the crossbar paradigm with another molecular nanotechnology for computation [11], [12]. While such solutions may be achievable in the long term, they require the integration of *two* different molecular nanotechnologies on the same surface which significantly increases manufacturing complexity. Keeping in mind the difficulties of fabricating a molecular crossbar technology by itself, integrating such a crossbar on top of a prefabricated CMOS IC, as shown in Fig. 1, will be easier to achieve, yet provide a robust computing paradigm. We refer to this design paradigm as Nano on CMOS (NoC), similar ideas being also suggested in [2].

The NoC paradigm also allows for significant design versatility. For example, while the nano portion is restricted to regular structures, the CMOS portion can be any arbitrary circuit. A number of design scenarios can be envisioned depending on the physical characteristics of the nano. One extreme is with the CMOS as the primary computation medium while the nano on top is used as a supplement to better achieve integration goals. For example, the nano crossbar could act as memory or large logic arrays. Likewise, at the other extreme, the nano portion would be primary while the underlying CMOS would be used simply to provide signal gain and latching capabilities. A more balanced approach uses both mediums for primary computation with portions of the circuit being allocated either to CMOS or nano at a finer grain. In the next section we address partitioning as well as an overall design methodology for such a mixed CMOS/nano approach.

III. CMOS/nano Co-design

The general concept of a mixed CMOS/nano circuit is to divide the functionality between a conventional CMOS technology and a nanoelectronic technology (nano). We consider a bottom-up nanotechnology that is restricted to regular circuit structures. Other characteristics dependent on the specific nanotechnology, such as, switching speed, area, power, and defect densities will also play a role in the partitioning process.

New design methodologies are needed for mixed CMOS/nano circuits. The possibility of high device densities for nano combined with the present challenges of CMOS design point towards a highly automated methodology. Fig. 2 shows a generic design methodology for a CMOS/nano circuit. This generic methodology is an adapted version of a typical ASIC design methodology and targets scenarios where portions of the circuit can be allocated to either CMOS or nano at a fine grain. The key feature in the figure is the partitioning that occurs after RTL synthesis. Fig. 2 also shows our proposed partitioning procedure in expanded detail. The partitioning procedure requires information about the CMOS process characteristics, such as a high level description of the standard cell library, including gate delay, area, and power estimates for the cells. The nano process also needs to be characterized at high level. For example, the crossbar technology described in the previous section cannot implement sequential logic or produce signal gain. Mechanisms for determining the maximum crossbar dimensions need to be supplied in the case of a technology lacking signal gain, as we show in the next section. In addition to functionality, delay, area, power, and defect densities also need to be included in the nanotechnology characterization. Another important metric is the overhead associated with interfacing the CMOS and nano portions of the circuit. We discuss CMOS/nano interfacing in more detail in the next section. The lack of gain in the crossbar technology requires periodic interfacing to CMOS circuitry to restore signal integrity. Signal restoration circuitry consists of the equivalent of a sense amplifier and a buffer to drive the next crossbar. Thus switching design mediums and restoring nano signals will come with an overhead. Using the CMOS/nano technology characterization and the logic level representation obtained from synthesis, the partitioning procedure goes through four phases of allocation:

Pass 1 *Default Allocation* - This pass allocates to CMOS the portions of the circuits that cannot be implemented in nano. Continuing with the crossbar example, this would include all analog and amplification portions of the design.

Pass 2 *Global Allocation* - Taking the design constraints and objectives as an input, this pass allocates the portions of the design that are inherently suitable to one of the technologies. For example, the critical path of the circuit may be allocating to the technology that can perform faster computation. The HP/UCLA crossbar technology is predicted to have slower switching speeds than CMOS, thus the critical path of the circuit will be implemented in CMOS. On the other hand, the HP/UCLA technology is predicted to consume less power than CMOS, so large regular structures, such as RAM, will be suited for implementation in nano. In general, it is expected that the I/O, BIST, control logic, and sequential processing is better suited for CMOS implementation, while parallel processing and memory

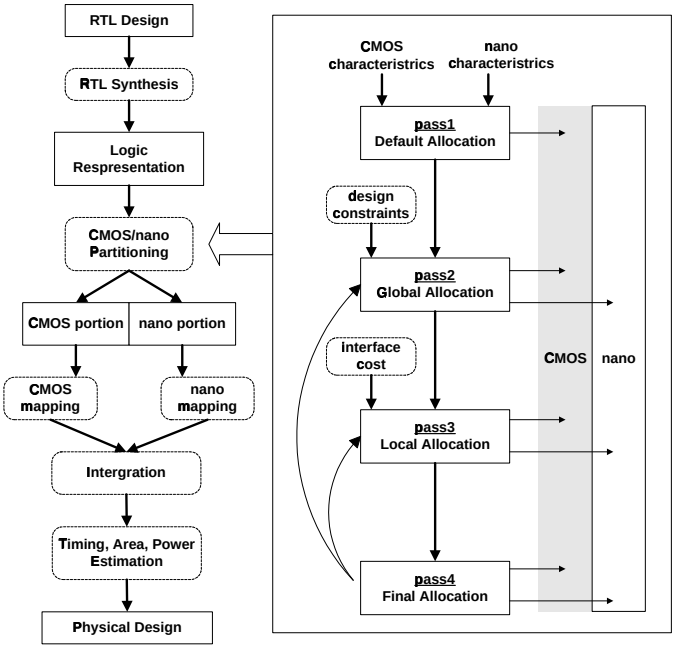


Fig. 2. Generic CMOS/nano design methodology (adapted ASIC methodology) and expanded allocation procedure between CMOS and nano.

Sum cubes	Cout cubes
CinAB	CinAB
	Cin¬AB
	CinA¬B
	¬CinAB
Cin¬A¬B	
¬CinA¬B	
¬Cin¬AB	

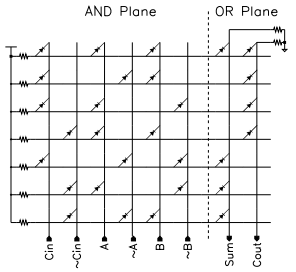


Fig. 3. A PLA representation of a full-adder mapped to a nano crossbar technology employing bistable diode junctions.

is more suitable for nano implementation.

Pass 3 *Local Allocation* - This pass uses the interface cost as an aid to determine if circuitry near the previously allocated portions should be located on the same medium.

Pass 4 *Final Allocation* - The remaining portions of the circuit are allocated to the optimal medium.

A brief example of partitioning for a generic circuit would be to allocate I/O, BIST, analog circuitry, control logic, and sequential logic to CMOS. In turn, disk storage, main memory, and parallel processing would be allocated to nano. Cache would be allocated to nano if access times are faster enough, otherwise it would be realized in CMOS. The remaining logic would be allocated to the appropriate medium, taking design constraints and interface costs into consideration.

While the first allocation pass occurs only once, the second through fourth allocation passes can iterate until an optimal solution is located. Following the partitioning procedure, the generic CMOS/nano co-design methodology performs technology mapping. The CMOS portion of the design follows a typical ASIC technology mapping flow. On the other hand, the nano portion uses a different method for technology mapping, which will be discussed in the next section.

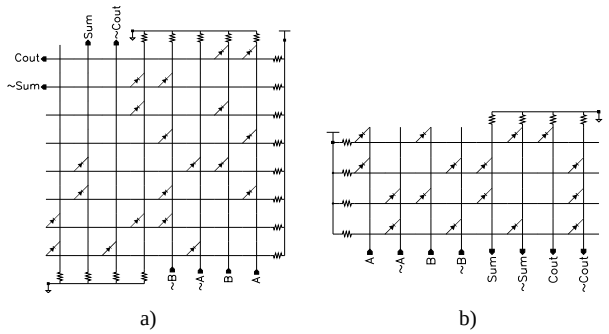


Fig. 4. a) A half-adder mapped using an arbitrary logic representation, b) A half-adder mapped using a two-level logic representation (PLA).

IV. Nano Backend Analysis

It will be necessary to supply ample backend information to earlier steps of the CMOS/nano co-design cycle, such as synthesis and partitioning. The ability to quickly evaluate potential tradeoffs for nano will allow better decisions early in the design cycle. In this section we analyze a nanotechnology based on arrays of crossed nanowires and present approaches that can be used for fast, yet accurate, estimations.

Relying on self-assembly for fabrication restricts the nano portion of the design to regular structures. The product of self-assembly is typically a blank fabric that requires programming for unique functionality. Regular structures are inherently suited for memory, while LUTs and PLAs are regular structures that can be used for implementing logic. For example, Fig. 3 shows a full-adder mapped to a crossbar with diode-like junctions like the HP/UCLA crossbar paradigm. There are two scenarios to be considered in terms of technology mapping. One scenario involves a self-assembly process that can produce only uniformly sized crossbar arrays. This scenario makes technology mapping for a molecular crossbar technology similar to an FPGA. A second scenario allows the dimensions of each crossbar array to be different and controlled at fabrication time. One method of controlling the coarse dimensions of a self-assembly process is to use lithography to define the borders, sometimes referred to as directed assembly [13]. This second scenario allows for optimizing results. Optimization in a flexible crossbar paradigm involves three factors: the optimal logic mapping, the cost of interfacing to CMOS, and the size restriction on the array due signal degradation.

A. Optimal Logic Mapping

One aspect of optimal logic mapping consists of choosing the appropriate logic representation, such as multi-level logic (arbitrary structures), minimized two-level representation (PLA structures), or the minterm canonical logic representation (LUT structures). The minimized two-level representation is the most appropriate for crossbar structures that can easily implement both PLAs and LUTs. Fig. 4 compares a half-adder implemented both from the a) PLA representation and b) arbitrary logic (reconstructed from [11]). This example, as well as other test circuits, favors PLA implementation in terms of array size. When considering a crossbar circuit, area is a particularly important metric because it dictates the interconnect capacitance and resistance, which in turn influences other metrics, such as delay and power. However, a possibly greater advantage of a PLA implementation is that the logic level representation is directly proportional to the physical array size, which allows for fast esti-

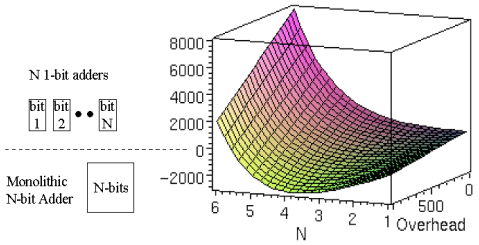


Fig. 5. The area consumed by an N-bit ripple-carry adder allocated to a single crossbar versus allocating a 1-bit adder to N crossbars.

mates of the physical design early in the design cycle. Choosing a PLA representation over a LUT representation essentially involves trading a universal logic array for a more optimal logic array. Equations (1) and (2) govern the size of a diode-based crossbar, for LUT and PLA structures similar to the circuit in Fig. 3. In the equations, N is the number of literals in all the functions implemented on the crossbar, f is the number of functions, and c is the number of two-level minimized cubes in all the functions. Equation (3) shows the area overhead for a LUT structure versus a PLA structure. Thus, the optimality of a PLA representation increases as more functions with overlapping product terms are allocated to a single crossbar.

$$LUT_{area} = 2^N(2N + f){P_{wire}}^2 \quad (1)$$

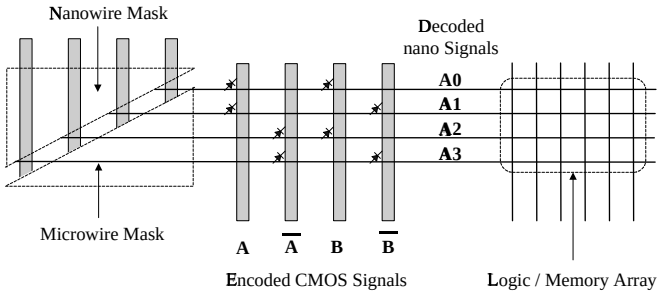
$$PLA_{area} = c(2N + f){P_{wire}}^2 \quad (2)$$

$$PLA_{savings} = 2^N/c \quad (3)$$

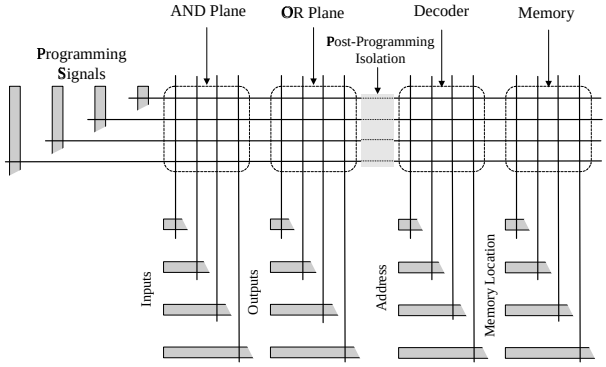
B. Interface Cost

There will be some overhead incurred when a signal switches mediums. Furthermore, the lack of signal gain in some crossbar technologies mandates that the computation must leave the crossbar periodically for restoring signal integrity. CMOS signal restoration will consume area, power, and cause signal delay. However, since very large PLA structures become inefficient, a network of nano PLAs is more desirable. Given the potential interface cost associated with leaving a crossbar, an important design decision involves whether to allocate many functions to a few large crossbars or fewer functions to a many smaller crossbars. Using a PLA logic representation we can estimate the tradeoffs involved in these decisions at a high level. One way to measure the size of a crossbar is to add together the unit-crossbar areas. A unit-crossbar area is simply the pitch of a horizontal wire times the pitch of a vertical wire. We analytically explore tradeoffs in crossbar granularity by representing the interface overhead in terms of unit-crossbar areas. Fig. 5 shows an example of these tradeoffs associated with an N-bit adder. The comparison involves an N-bit carry-ripple adder implemented in a single crossbar compared to a chain of 1-bit full-adders with an interface overhead incurred for each full adder. The vertical axis of the figure is the area of the single N-bit PLA adder substracted from the N 1-bit PLA adder area. Therefore, the design space with negative vertical-axis values favors a single PLA implementation, whereas the design space with positive vertical-axis values has an optimal multiple-PLA implementation. Likewise, compromises between these two extremes can be explored, such as implementing an N-bit adder on M PLAs, with $M < N$.

A more general interface problem involves the mismatch of CMOS wire pitches and nano wire pitches, which is complicated by the restriction to regular nano topologies. We will use the term microwires to refer to the wires in the CMOS portion of the design



a)

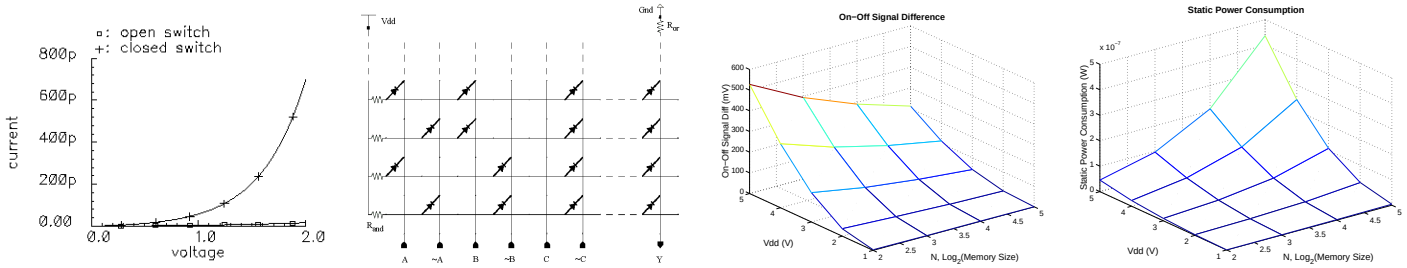


b)

Fig. 6. a) A CMOS/nano interface structure that relies on mask alignment precision and a programmed decoder, b) A second decoder design that illustrates integrated logic and memory.

and nanowires to refer to wires in the nano portion of the design. Interfacing microwires and nanowires end-to-end is not desirable because the device density of the nano crossbar would be limited by the CMOS process wire pitch. In addition, it is expected that the density of the nano portion is high enough to warrant addressing the nano crossbars in an encoded fashion. Thus, micro to nano decoders and encoders, also referred to as demuxes and muxes, have been proposed to interface the two technologies as well as provide a solution to the microwire and nanowire pitch mismatch [7], [12]. These designs rely on stochastic assembly and/or fabrication control of irregular features at nanoscale resolution.

To avoid these potential fabrication problems, we propose a new decoder design for interfacing CMOS and nano. Instead of requiring stochastic assembly and irregular nanoscale patterns, our designs relies on precision mask alignment and programming the decoder into the crossbar. Fig. 6 a) shows how the microwire pitch can be reduced to the nanowire pitch by using on-off masks aligned diagonally to produce a one-to-one microwire to nanowire correspondence. The decoder is then programmed into the crossbar after fabrication. The structure in Fig. 6 a) demonsrates how microwire address lines can be built into the decoder. Likewise, as shown in Fig. 6 b), the diagonally aligned masks can be used to reduce the pitch before the decoder. Fig. 6 b) also shows how multiple arrays can share the programming microwires with the aid of a post-programming isolation mechanism. A second nice feature of the structure in Fig. 6 b) is the ability to integrate logic and memory. While logic is programmed using the microwires following fabrication and only read thereon after, memory requires the ability to read and write data multiple times using the decoder. To allow data to be written to the memory array without overwriting the programmed junctions in the de-



a) b) c) d)

Fig. 7. Simulation of a circuit employing a programmed decoder realized in a diode-based crossbar technology: a) Crossbar junction I-V curves, b) General crossbar schematic, c) On-off signal difference, c) Static power consumption.

coder, we suggest employing a decoder composed of junctions that are programmed at different voltages than the memory array junctions. We believe this is feasable because a nanoscale pitch is not required between the decoder and memory array. Thus, a nanowire array could be placed at a lithograhically defined distance from an another nanowire array consisting of a second type of devices, e.g., dissimilar arrays could be made by using different molecular junctions or growing oxides of different thicknesses for each array. The decoder schemes in Fig. 6 are also beneficial in terms of defect tolerance. Since the arrays begin as blank fabrics, defective wires can be located through testing and signals can be re-routed to spare wires. As in [11], this defect tolerance approach is more effective for faults resulting in open circuits than for faults causing closed circuits.

C. Physical Limitations

Although analysis at higher levels of abstraction may suggest certain optimal crossbar mappings, the physical size of the crossbar arrays will be limited by the characteristics of the crossbar technology as well as the design goals and constraints.

To explore the design space of the programmed decoder we introduced previously, we model the bistable rotaxane crossbar junctions from [3] and [4] using Verilog-A, as shown in Fig. 7 a). We simulate the decoder reading a memory array for variable array sizes using Cadence's Spectre circuit simulator. Fig. 7 b) shows a generalized schematic. The explicit resistors R_{and} and R_{or} are set to near optimal values and the memory array junction diodes are programmed to bring out worst case scenarios. The lack of gain in the diode-based crossbar technology as well as the non-ideal behavior of a diode-based decoder, i.e., leakage by diode-resistor AND gates, cause the worst-case on/off voltage difference to decrease as the size of the decoder and the memory array grow. One mechanism to increase the worst-case on/off voltage is to increase Vdd. Fig. 7 c) shows our simulation results for worst-case on/off voltage as the size the of the array and Vdd are varied. However, an increase in Vdd will in turn raise the power consumption, as shown in Fig. 7 d). Thus, the resolution of CMOS sense amplifiers may restrict the sizes of crossbar arrays and power consumptions constraints may resrict raising Vdd. Crossbars operating at higher supply voltages will also require level shifters to convert between CMOS voltage levels and nano voltages. Furthermore, the behavior of the rotaxane junctions needs to be reconsidered for larger supply voltages. Another issue involves whether the heat dissipated by the CMOS circuitry will effect the the nano crossbars on top. But in general, seeing that the nano crossbar paradigm is still at an early stage of development, there remains a variety of ways for improvement, such as employing diodes with more ideal characteristics or the incorporating other devices, such as, nanoscale transistors.

V. Conclusion

Extending Moore's Law past silicon's physical and economic barriers will require new nanoelectronic solutions. However, the benefits of silicon integrated circuits will present difficult competition for novel up-and-coming nanotechnologies. We propose an approach stressing "peaceful coexistence" between silicon and a partner nanotechnology. Mixed CMOS and nano circuits will require new fabrication and design paradigms. In this paper we have presented a methodology for CMOS/nano co-design. In addition, we have also considered the logical and physical aspects of a nano crossbar technology. While nanoelectronics is still quite young, combining proven CMOS design strategies with new novel nanoelectronic design approaches can lead to higher levels of computation.

VI. Acknowledgments

Thanks to Dr. James Ellenbogen from the MITRE Corp. for interesting discussions on this topic.

References

[1] "International technology roadmap for semiconductors 2001 edition," .

[2] "Technology roadmap for nanoelectronics 2000 edition. http://www.cordis.lu/esprit/src/melna-rm.htm," .

[3] C. P. Collier, E. W. Wong, M. Belohradsky, F. M. Raymo, J. F. Stoddart, P. J. Kuekes, R. S. Williams, and J. R. Heath, "Electronically configurable molecular-based logic gates," *Science*, vol. 285, pp. 391–394, July 1999.

[4] P. J. Kuekes, J. R. Heath, and R. S. Williams, "Molecular wire crossbar memory," US Patent, number 6128214 (Hewlett-Packard), October 2000.

[5] Y. Huang, X. Duan, Y. Cui, L. J. Lauhon, K. Kim, and C. M. Lieber, "Logic gates and computation from assembled nanowire building blocks," *Science*, vol. 294, pp. 1313–1316, November 2001.

[6] P. J. Kuekes, J. R. Heath, and R. S. Williams, "Molecular-wire crossbar interconnect (MWCI) for signal routing and communications," US Patent, number 6314019 (Hewlett-Packard), November 2001.

[7] P. J. Kuekes and R. S. Williams, "Demultimplexer for a molecular wire crossbar network (MWCN DEMUX)," US Patent, number 6256767 (Hewlett-Packard), July 2001.

[8] A. R. Pease, J. O. Jeppesen, J. F. Stoddart, Y. Luo, C. P. Collier, and J. R. Heath, "Switching devices based on interlocked molecules," *Acc. Chem Res.*, vol. 34, no. 6, pp. 433–444, 2001.

[9] Y. Cui and C. M. Lieber, "Functional nanoscale electronic devices assembled using silicon nanowire building blocks," *Science*, vol. 291, pp. 851–853, February 2001.

[10] W. J. Gallagher, J. H. Kaufman, S. S. Papworth, and R. E. Scheuerlein, "Magnetic memory array using magnetic tunnel junction devices in the memory cells," US Patent, number 5640343 (IBM), June 1997.

[11] S. C. Goldstein and M. Budiu, "Nanofabrics: Spatial computing using molecular nanoelectronics," in *28th International Symposium on Computer Architecture*, June 2001.

[12] A. DeHon, "Array-based architecture for molecular electronics," in *1st Workshop on Non-Silicon Computation (NSC-1)*, 2002.

[13] Y. Huang, X. Duan, Q. Wei, and C. M. Lieber, "Directed assembly of one-dimensional nanostructures into functional networks," *Science*, vol. 291, pp. 630–633, January 2001.