By Warren Robinett, Gregory S. Snider, Philip J. Kuekes, and R. Stanley Williams

COMPUTING WITH A TRILLION CRUMMY COMPONENTS

Attempting to build nanometer-scale circuits that are both defect- and fault-tolerant.

In the 1950s, computers were built using a variety of components, including vacuum tubes and relays. These devices had the tendency to burn out during a computation, which was a significant disruption since the operator had to physically replace the component and restart the computation. There was also a high probability a device would fail intermittently, which meant the computations had to be performed multiple times and the results compared in order to assure the results were correct. Thus, computer scientists of that time, including John von Neumann [12] and Claude Shannon [6][1], began to seriously examine the possibility of building machines that could operate perfectly even if their components were defective or unreliable. This pioneering work slipped into obscurity in the 1960s with the advent of integrated circuits, which were so trustworthy they were essentially flawless. However, after 40 years of refinement, the dramatic shrinkage of the device sizes in integrated circuits is reaching a point where such quality will no longer be possible. We will soon be back to the point at which we will want to know how to build absolutely reliable systems with crummy [6] components.

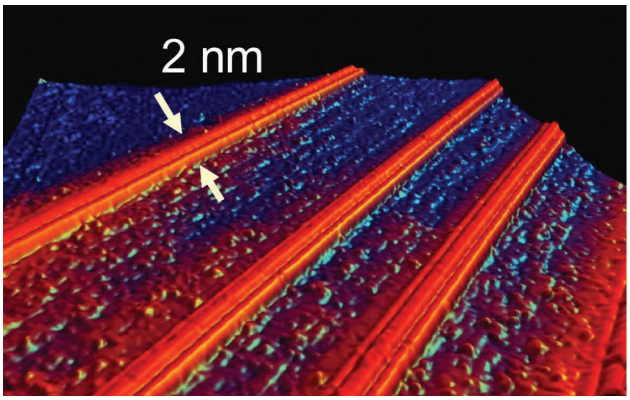
[1]Moore and Shannon provided a technical definition of the word "crummy" with respect to a relay based on the probabilities that it would fail to open or close properly when required.

At this stage, it is not even certain what those components will be. At size scales less than 10 nanometers (which is equivalent to approximately 40 silicon-silicon atomic bond lengths—see the sidebar "The Nanometer Challenge"), the operation of transistors will be highly problematic. It is possible and even likely that new types of switching devices will be utilized in future circuits, either in combination with some transistors or all by themselves. These switches will have very different operating characteristics from standard silicon devices, and they may have operating characteristics similar to the relays used in the 1950s—those that are being studied today are built from small clusters of molecules or a very thin layer of an oxide material between two metal electrodes. They are essentially nanometer-sized electrochemical cells—similar to batteries but smaller than a virus—that can be toggled open (high resistance) or closed (low resistance) by placing a potential across the device that exceeds a threshold voltage. The opening and closing threshold voltages usually have opposite polarities that drive chemical reactions involving only a few molecules or atoms in the switch, but which change the electrical resistance of the device dramatically by four or more orders of magnitude. The new components, in this quantity and at this scale, will probably work as desired, but some will not work at all and others will fail at random intervals. We call the completely broken components defects, and the intermittent mistakes faults. Our goal is to build nanometer-scale circuits that are both defect- and fault-tolerant [3], since we cannot replace broken devices and we prefer not to rerun a computation unless absolutely necessary.

Switches can be utilized as the basis for a memory or a logic circuit. The trick is wiring up a huge number of them to perform a useful task. Since we anticipate that these switches will eventually be less than 10 nanometers (nm) wide, one trillion (10^{12}) of them will fit onto a one-square-centimeter chip surface. Thus, we will require a large number of very small wires to connect up all of these switches. The simplest architecture to accomplish this task is the crossbar [5], which is a very familiar structure in various types of networks. To connect 10^{12} switches in a single crossbar, one would have one million parallel wires spaced 10nm from center to center on the bottom crossed over by another million wires at right angles to the first set, with a switch at the intersection of each pair of crossing

THE NANOMETER CHALLENGE

In the future, it may be more convenient to allow selected circuit components to form themselves rather than fabricating them by using some type of lithography. This image shows a set of parallel wires that are 2nm wide and separated on average by 9nm, just the size that we need to make our trillion-device crossbar. In order to see the wires at all requires a special experimental probe called a scanning tunneling microscope (STM), which collects data by raster scanning an extremely sharp needle over a sample and measuring the flow of electrons, called a tunneling current, from the needle to the sample. The measured current vs. position data are turned into a visual image using computer graphics. This technique is so sensitive that one can observe individual atoms, which are evident as the small bumps in the image. These wires are only six atoms wide, and they were formed by a process called self-assembly during a chemical reaction between a thin film of vapor-deposited erbium atoms and the atoms on the surface of a standard silicon wafer. This image demonstrates that at this scale, matter is not smooth but really is made of discrete atoms. Just one or two atoms out of place will create a significant defect in a wire.



A set of 2nm-wide parallel wires.

Thermodynamics guarantees such defects will occur, and occur fairly frequently. Just as we have for some time had Einstein the traffic cop telling us our signals cannot exceed the speed of light, we now also have Boltzmann the construction foreman waving the Second Law of Thermodynamics at us, saying we will never get every atom in the desired place. We must accept these fundamental physical limits. But just as houses get built imperfectly, yet rarely fall apart, we can "overengineer" our nanoscale computers to tolerate imperfection. Thus, a new set of ground rules for computer design is emerging, and this will provide a new set of challenges for the forthcoming generation. C

wires. However, the expense of ensuring that all of the switches will operate perfectly would be astronomical, so in order to keep manufacturing costs reasonable, a significant fraction (estimated to be ~10% from prototype circuits we have built) of the devices will be nonfunctioning at manufacturing time. Even if the operational devices have a mean time to failure of 100 years, that will still mean the rate of component failure will be one every seven milliseconds if the distribution is constant. Since it is impossible to replace or repair any of these nanoscale components, the need for reliability will be met through redundancy of the nanowires and switches [4]. Applied cleverly, a relatively small amount of redundancy can provide a substantial amount of protection from defects and faults [7].

In order to build and evaluate actual nanometer-scale circuits, we utilize a process called nano-imprint lithography (NIL) to make the wires and switches, as shown in Figure 1 [2]. This is an inexpensive process that can be used to make extremely small features in a research setting. At the present time, we can make nanowires that are approximately 15nm wide, whereas the smallest wires in today's most aggressive commercially available semiconductor circuits are ~65nm wide.

Now we have the challenge of assembling switches on these wires. If we were to attempt to place switches down one at a time in perfect ordered rows, the cost and the time required for doing so would be enormous. Therefore, we use a self-assembling and self-aligning technique, in which we cover the entire substrate, wires and all, with a uniform thin layer of a switching material. This material can be applied as a monolayer of molecules or as a thin film of an oxide, for instance. Then, a second layer of nanowires is formed on top of the switching layer by NIL. The final step is to use a chemical etching process to remove all of the switching material that is not directly under the top set of nanowires, which acts to isolate all of the junctions from each other. The final crossbar structure is shown in Figure 2. Every nanowire in the top layer is connected to a nanowire in the bottom layer through a switch, which can either be set open (a high-resistance connection) or closed (a low-resistance connection).

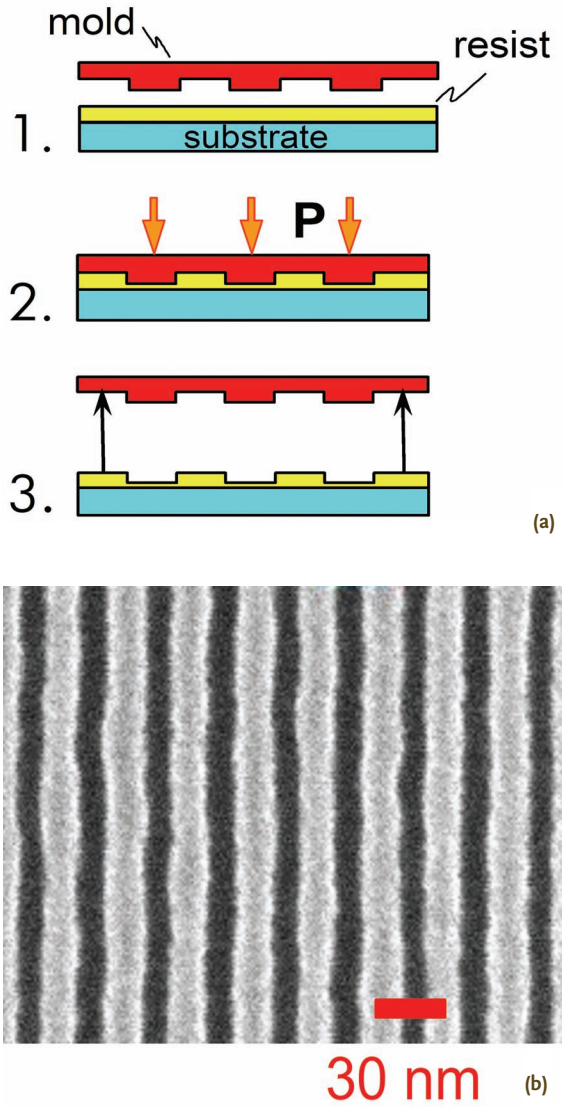


Figure 1. (a) Nano-imprint lithography procedure. (b) A mold is pressed into a thin layer of polymer that coats a silicon substrate and is then removed, leaving the wire-array pattern. This is transformed into a set of parallel metal wires by a subsequent sequence of chemical etching and deposition steps. (b) Scanning-electron microscope (SEM) image of metal wires with a center-to-center spacing of 30nm made by NIL.

It is fairly easy to see how such a structure can be used for a random access memory [11]. If we choose a switch-open state to represent a '0' and a switch-closed state to represent a '1', it is possible to write 0s or 1s into the crossbar just by applying the appropriate voltage across a pair of wires that has an active switch between them. We can later measure the resistance between those two wires in order to read out the data value stored at that location. In the case of memories, defect and fault tolerance can be achieved by designing circuits that incorporate concepts from coding theory developed by Shannon for sending messages through noisy environments [1].

We have developed several new logic families that utilize switches for computing. For this article, we choose as an example a hybrid technology that combines both nanoscale switches and microscale CMOS transistors to dramatically improve the performance of a particular type of computing circuit—a field-programmable gate array (FPGA)—as a means of capitalizing on the dense crossbar fabric.

An FPGA chip is a piece of programmable hardware, comprising a set of gate-level building blocks, such as NAND gates and flipflops, and a data-routing network that can semi-permanently wire together these components into a digital circuit. The wiring pattern is implemented with CMOS switches controlled by configuration bits, stored in static memory cells. For a hardware engineer, there is often a trade-off between the high performance (and high cost and long design time) of a custom ASIC chip versus the lower performance (and lower cost and immediate availability) of using a general-purpose processor; an FPGA is halfway between these two alternatives, and may be described as semi-custom hardware.

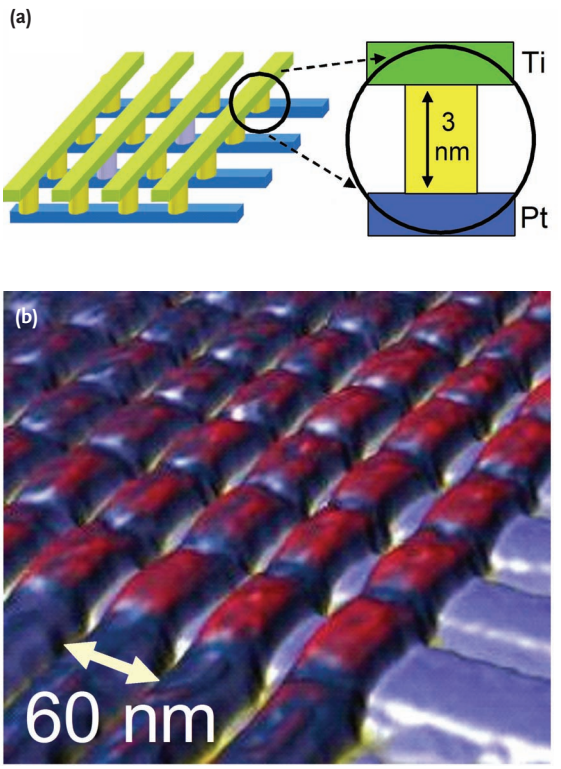


Figure 2. (a) The crossbar geometry. Each junction where a titanium nanowire (green) crosses over a platinum nanowire (blue) has a thin layer of active switching material (yellow) between the wires. (b) An atomic force microscope image of a section of an actual 34 x 34 crossbar fabricated in our research group.

Our hybrid technology is called field-programmable nanowire interconnect (FPNI) [8–10]. The basic idea is to make a hybrid nanoswitch/CMOS chip, using the switch components only for configurable interconnect, and using standard CMOS microcircuitry for all other functions, such as logic and configuration of the interconnect (see Figure 3). Compared to the standard FPGA architecture, this approach takes all of the CMOS resources required for the configuration bits and switches and replaces them with a set of nonvolatile nanoswitches residing in the metal interconnect layer.

Since nanoswitches are much smaller than the configuration bits that consume a large fraction of the area in a traditional FPGA, this design offers much greater logic density. In a recent study [8], we describe a FPNI circuit with a factor of eight increase in logic density, comparable clock speed and reduced power dissipation compared to a CMOS-only FPGA, using the same transistor technology. This improvement is equivalent to three generations of CMOS development, or nearly 10 years equivalent to Moore's Law technological progress, without having to shrink or improve the transistors in the circuit. Thus, we see this example as an existence proof that the performance of CMOS technology can be extended well beyond currently extrapolated limits by optimizing the metal interconnect, which is already a major performance limitation in terms of operating speed and power dissipation.

The dense interconnectivity, high bandwidth, uniformity, and sparse utilization of the nanowire crossbar enables very effective schemes for defect tolerance [3]. This is essential, since high defect rates are inevitable for any technology that incorporates nanoscale devices, independent of their composition or function. Our simulations of a defect-avoidance strategy [8] in the FPNI chip showed that even at a defect rate of 50% for the nano junctions, the effective manufacturing yield was 99.7% with little degradation of the circuit speed. Thus, we believe that it is possible to introduce hybrid technologies,

THIS IMPROVEMENT IS EQUIVALENT TO THREE GENERATIONS OF CMOS DEVELOPMENT, OR NEARLY 10 YEARS EQUIVALENT TO MOORE'S LAW TECHNOLOGICAL PROGRESS, WITHOUT HAVING TO SHRINK OR IMPROVE THE TRANSISTORS IN THE CIRCUIT.

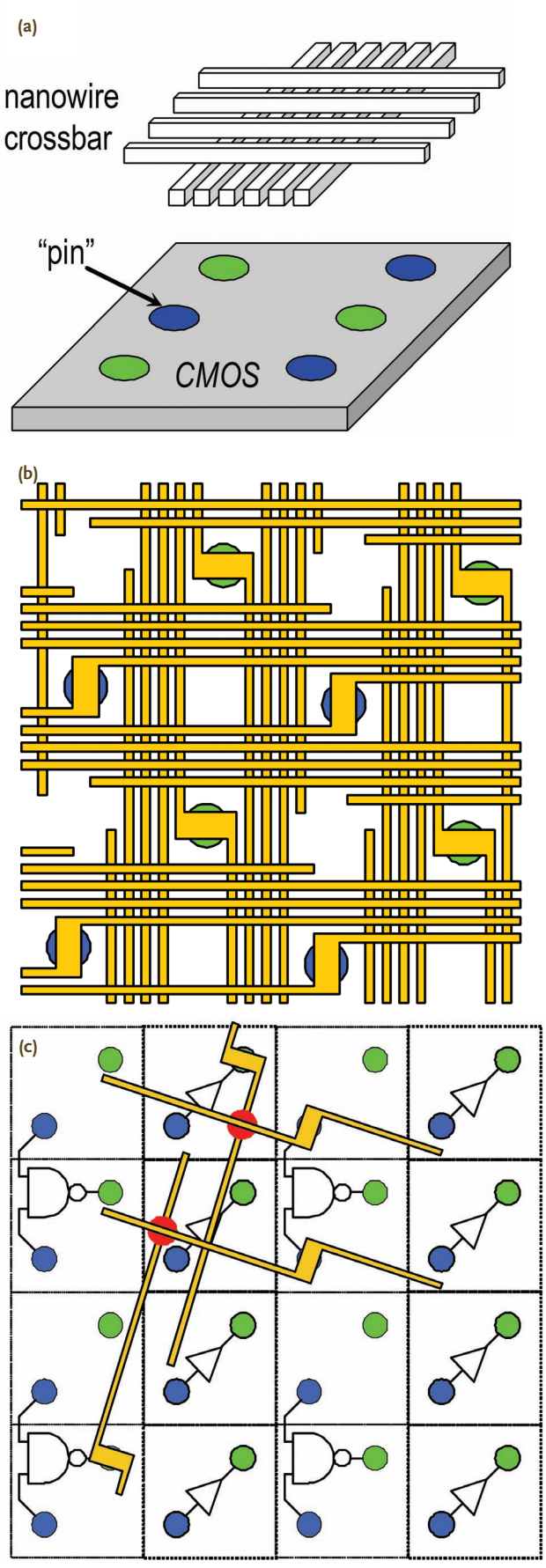


Figure 3. (a) The nano-scale crossbar and the CMOS micro-circuitry are fabricated as separate layers of the chip, with an "area-distributed" electrical interface (blue and green pins) between the two layers [10]. (b) Green pins connect vertical nanowires in the nano layer to gate inputs in the CMOS layer; blue pins connect horizontal nanowires in the nano layer to gate outputs in the CMOS layer. (c) The nanowire junctions are open by default, and by closing selected nano-switches (red dots) to connect certain horizontal and vertical nanowires, signal paths are formed to connect gate outputs to inputs. This is equivalent to the configuration done in traditional FPGA chips. Two configured data paths are shown.

in which some type of nanoscale switch is used to complement CMOS, and thereby continue Moore's Law rates of improvement in computing capacity while maintaining reliability for many more decades into the future with only modest improvements in transistors. C

References

1. Chakraborty, K. and Mazumder, P. *Fault Tolerance and Reliability Techniques for High-Density Random-Access Memories*. Prentice Hall, 2002.
2. Chen, Y. et al. Nanoscale molecular switch devices fabricated by imprint lithography. *Applied Physics Letters 82* (2003), 1610–1612.
3. Heath, J.R., Kuekes, P.J., Snider, G.S., and Williams, R.S. A defect-tolerant computer architecture: Opportunities for nanotechnology. *Science 280* (1998), 1716.
4. Kuekes, P.J., Robinett, W., Seroussi, G., and Williams R.S. Defect-tolerant interconnect to nanoelectronic circuits: Internally-redundant demultiplexers based on error-correcting codes. *Nanotechnology 16* (2005), 869.
5. Kuekes, P.J. and Williams, R.S. Demultiplexer for a molecular wire crossbar network. U.S. Patent No. 6,256,767 (July 3, 2001).
6. Moore, E.F. and Shannon, C.E. Reliable circuits using less reliable relays. *Journal of the Franklin Institute* (1956), 191–208, 281–297.
7. Siewiorek, D.P. and Swarz, D.S. *Reliable Computer Systems: Design and Evaluation, 3rd edition*. AK Peters, Ltd., 1998.
8. Snider, G.S. and Williams, R.S. Nano/CMOS architectures using field-programmable nanowire interconnect. *Nanotechnology*, 2006.
9. Strukov, D.B. and Likharev, K.K. CMOL FPGA: A cell-based, reconfigurable architecture for hybrid digital circuits using two-terminal nanodevices. *Nanotechnology*; preprint available at rsfq1.physics.sunysb.edu/likharev/nano/FPGA05.pdf.
10. Strukov, D.B. and Likharev, K.K. A reconfigurable architecture for hybrid CMOS/nanodevice circuits. In *Proceedings of FCCM'05* (Napa Valley, CA, April 2005); preprint available at rsfq1.physics.sunysb.edu/likharev/nano/FCCM2005.pdf.
11. Strukov, D.B. and Likharev, K.K. Prospects for terabit-scale nanoelectronics memories. *Nanotechnology 16*, 137 (2005).
12. von Neumann, J. Probabilistic logics and the synthesis of reliable organisms from unreliable components. In C.E. Shannon and J. McCarthy, Eds., *Automata Studies* (1955), 43–98.

Warren Robinett (warren.robinett@hp.com) is a computer architect with the Quantum Sciences Research Group at Hewlett-Packard Labs in Palo Alto, CA.

Gregory S. Snider (snider.greg@hp.com) is a computer architect with the Quantum Sciences Research Group at Hewlett-Packard Labs in Palo Alto, CA.

Philip J. Kuekes (philip.kuekes@hp.com) is a computer architect with the Quantum Sciences Research Group at Hewlett-Packard Labs in Palo Alto, CA.

R. Stanley Williams (stan.williams@hp.com) is a senior fellow with the Quantum Sciences Research Group at Hewlett-Packard Labs in Palo Alto, CA.

© 2007 ACM 0001-0782/07/0900 $5.00