The Quest for the Next Information Processing Technology

Jeffrey J. Welser
Semiconductor Research Corporation
Nanoelectronics Research Initiative
IBM Almaden Research Center

In recent semiconductor technology generations, exponentially increasing power density has started to limit the historical benefits of scaling. Entirely new device approaches and methods of computation in emerging nanoscale technologies need to be explored. The Nanoelectronics Research Initiative (NRI) is taking on the grand challenge to find a "new switch" which can continue the annual exponential increase in information processing capability which has benefited not only the semiconductor industry, but nearly every aspect of our electronics and information technology driven modern economy.

INTRODUCTION

For over three decades, the semiconductor industry has been driven by its ability to scale the size of the Complimentary Metal Oxide Semiconductor (CMOS) Field-Effect Transistor (FET), the key building block in modern integrated circuit (IC) chips. This scaling has enabled the industry to pack twice as many FETs onto a chip every 18 – 24 months, in what has come to be known as "Moore's Law" (Moore-1965). The result has been an exponential increase in the information processing capability per unit area on the chip – or more importantly, per dollar.

Recently, exponentially increasing power density, due to both FET leakage currents as well as switching energy, has been limiting the continuation of this scaling trend. The fundamental physics of FET operation, rather than fabrication capability, will likely be the ultimate limit for future scaling in the next 10-15 years. This has led to a quest for new devices that can continue the historical trends in information processing performance.

To take on this grand challenge, the Nanoelectronics Research Initiative (NRI) (nri.src.org) was formed in 2004 as a consortium of Semiconductor Industry Association (SIA) (www.sia-

online.org) companies to manage a university-based research program as part of the Semiconductor Research Corporation (SRC) (www.src.org). The NRI was founded by six U.S. semiconductor companies (AMD, Freescale, IBM, Intel, Micron, and TI), and partners with the National Science Foundation (NSF), the National Institute of Standards and Technology (NIST), and state governments, sponsoring research currently at 35 U.S. universities in 20 states.

NRI Mission: Demonstrate novel computing devices capable of replacing the CMOS FET as a logic switch in the 2020 timeframe.

- These devices should show significant advantage over ultimate FETs in power, performance, density, and/or cost to enable the semiconductor industry to extend the historical cost and performance trends for information technology.
- To meet these goals, NRI is focused primarily on research on devices utilizing new computational state variables beyond electronic charge. In addition, NRI is interested in new interconnect technologies and novel circuits and architectures, including non-equilibrium systems, for exploiting these devices, as well as improved nanoscale thermal management and novel materials and fabrication methods for these structures and circuits.
- Finally, it is desirable that these technologies be capable of integrating with CMOS, to allow exploitation of their potentially complementary functionality in heterogeneous systems and to enable a smooth transition to a new scaling path.

PHYSICS OF A LOGIC SWITCH

As outlined in the 1970's (Dennard-1974), if one shrinks the critical dimensions of an FET by a factor kappa, while simultaneously increasing the doping levels and decreasing the applied voltages by the same factor, the scaled transistor switches faster, but consumes half the power and takes up half the area. This means twice as many transistors can fit in the same area while power density remains constant – the primary reason scaling works without melting the chip. Leakage currents, however, increase as dimensions shrink. This includes leakage along the transistor channel when the switch is turned off and across the gate insulator, which has become so thin (<1.5nm or just a few atomic layers) that quantum tunneling dominates. The leakage power is now becoming equivalent to the active switching power of the transistor.

The 2001 ITRS Emerging Research Device Technical Working Group did a highly-simplified analysis of a generic electronic switch at thermal equilibrium (Zhirnov-2003). The switch was modeled as a potential barrier separating two quantum wells, corresponding to the simplest version of an FET channel between source and drain contacts. The analysis showed the channel could conceivably be scaled down to ~1.5 nm and the transistor could have a minimum switching speed of ~40 fs – significantly smaller and faster than today's FETs of about 30nm channel length and ~1ps switching time. However, in order to avoid leakage over the barrier at room temperature, the voltage could not be scaled as rapidly as the physical dimensions, and the resulting power density for these switches at maximum packing density would be on the order of 1MW/cm^2 – orders of magnitude higher than the practical air-cooling limit of ~100W/cm^2. The theory does not consider what materials or structures are used – it is applicable to any switch that moves charge dissipatively across a potential barrier – leading to two implications: (1) Simply shrinking an FET to the far nanoscale will not necessarily continue to give the historical benefit of scaling, as the increasing power density will require trading off switching speed for packing density. (2) The existing Si FET roadmap is likely to be able to reach the minimum practical dimensions in the next 10-15 years (ITRS-2007), and while using new FET materials or geometries can yield improved performance, it will not alter the ultimate scaling limits.

A new "switch" for information processing is needed to significantly extend the scaling path. To define research directions for this quest, groups from industry, government and academia participated in workshops sponsored by SRC, NSF, and SIA (Cavin-2004, 2005, 2006). Thirteen research vectors were defined, and the top five comprise the NRI research program.

New Devices: Alternative computational state vectors

An FET device moves electrons dissipatively to charge (discharge) capacitors to represent a binary "1" ("0"). NRI research focuses on finding alternative ways to represent these states or information in general. Any physical property that can be placed into two or more distinguishable states could potentially be used to represent information. One example is to use the spin of an electron, with spin "up" representing "1" and spin "down" representing "0". Spin is already used successfully in memory and storage devices, so the challenge is how to inject, manipulate, and read-out the spin state of an electron or collection of electrons to build logic gates and circuits, and many different devices are being considered under the broad heading of

"spintronics" (Zutic-2004). Many other materials offer different potential states which could be exploited for logic, including ferroelectric, antiferroelectric, ferromagnetic, antiferromagnetic, ferrotoroidic, ferroelastic and ferrimagnetic materials (Eerenstein-2006). It is even possible to consider physical movement of atoms as a new state variable. While atoms are more massive than electrons, it would only be necessary to move them on the order of angstroms to cause large changes in the material (e.g. changing the dipole in a ferroelectric or changing barrier heights at an interface), so that the speed and energy could still be reasonable. And for nanoscale devices, more massive particles are less likely to lose their state by tunneling (Zhirnov-2008).

Much of the work on new state variables relies on developing new materials. Dilute magnetic semiconductors offer potential to introduce spin into semiconductors (Pearton-2004), while multi-ferroic materials, which could couple ferroelectric and ferromagnetic parameters, could be used to manipulate spins without magnetic fields (Eerenstein-2006). A recently discovered material of particular interest is graphene – a single mono-layer of graphite, with unique transport properties (Geim-2007). The two-dimensional honeycomb lattice of graphene gives rise to a conical band structure which leads to electrons behaving as massless Dirac fermions. Graphene could not only improve FET devices, due to its high carrier velocity, but could also enable new devices exploiting its unique physics. The pseudospin property (Min-2008) for example potentially enables a correlated shift of charge density between two graphene layers, which could lead to a new low-energy switch.

New Methods for Computation: Non-equilibrium systems

Operating an FET at room temperature requires energy barriers of sufficient height to maintain distinguishability between states, and this will be a factor for any other device as well. One approach to get around this problem would be to re-capture the computation energy, rather than allowing it to dissipate as heat. This is the goal of adiabatic or "reversible" computation (Bennett-1988). Another approach would be to do computation out of equilibrium. In the solid state, local distributions of carriers (or spins or other phase states) can be out of equilibrium with the ambient "temperature" for a period of time, before relaxing to the lattice temperature through phonon collisions or other coupling parameters. If the relaxation time is sufficiently long, the potential barriers could be lower, allowing state manipulation with lower switching energy.

It has not yet been experimentally proven that either approach can be used for computation with reduced energy dissipation, but out-of-equilibrium behavior is a primary motivation for considering alternative state variables for information processing in the first place. If an alternative state variable obeys the same Boltzman statistics as a dissipative electron-based system, the chance of it offering substantial energy advantages over FETs is small.

New Ways to Connect Devices: Non-charge data transfer

Any computation system requires connecting multiple devices and transferring information between them. For electron devices, charge is the natural carrier to use, but this drives much of the power consumption in modern IC's. An alternative device should transmit the new state variable – converting back to charge would negate any advantage of the new information token. For example, a spin device should transmit spin to the next device, such as through a spin-wave bus (Khitun-2007). The movement of a spin wave can be very low energy (Bernevig-2006), but if moving electrons is required to move the spin, the power advantage is lost. Similarly, if the device uses ferromagnetic or ferroelectric orientation, that should be transmitted through some low-energy magnetic or lattice interaction. Transmitting information short distances by coherent waves or collective effects is a promising approach to interconnecting devices, which favors architectures based on nearest-neighbor device coupling. It will likely still be necessary to convert to charge for cross-chip interconnects, and for coupling back out to the external world.

New Methods to Manage Heat: Nanoscale phonon engineering

Finding more energy-efficient ways to cool devices is a very active area of research across the industry, given its immediate importance for current CMOS chips. The focus within NRI is limited to looking at ways to control phonon flow for more efficient phonon extraction and manipulation in device structures. This is coupled to the non-equilibrium system work, since finding ways to lengthen the time the state is out-of-equilibrium with the thermal environment could be key to enabling low-energy computation. It may even be possible to utilize phonons themselves as the state variable (Wang-2007). Current research focuses on room temperature operation, given the large costs in energy required to go to lower temperatures. If exceptionally efficient cooling mechanisms were discovered however, that boundary condition could change.

New Fabrication Methods: Directed self-assembly of devices

Directed self-assembly combines traditional patterning with self-organizing systems to create nanostructures. The ability to cost-effectively fabricate nanoscale CMOS is a challenge for the industry, with the focus on self-assembly being to improve lithography for continuing scaling. NRI works only on self-assembly for directly creating new device structures, such as arrays of self-assembled magnetic dots for magnetic quantum cellular automata (MQCA) circuits (Liu-2008, Bernstein-2005). Increased focus on fabrication is likely in the future, once the state variable and architecture have been established.

SUMMARY

The grand challenge to find a device capable of extending information processing beyond the ultimate limits of CMOS technology is quite daunting, similar to the challenge faced in the 1940's when solid state transistors were developed to replace vacuum tubes. The current NRI program (Welser-2008) is largely focused on the first research vector – finding a new device – since that would more clearly define the directions for research on the other vectors. However, the study of alternative devices, data transport, thermal transport and manufacturability ultimately needs to be tightly integrated, to bridge from the basic science to a practicable information processing technology.

The CMOS FET is a very efficient switch, and many of the limits it is approaching in the next decade are fundamental to any device operating at room temperature. While it is difficult to predict what device might be capable of surmounting these limits, a few educated guesses can be made: To reduce power, it will likely be slower and rely on local interconnects. To compensate, it will need to be densely packed and likely utilize three-dimensional architecture. And cost-effective manufacturing will favor uniform arrays of devices – potentially self-assembled – with robustness at the device or architecture level to the increasing variability at the nanoscale. What information processing architecture is capable of utilizing such a device? The brain of course is often cited as a proof-of-concept that this can be done, at least for certain types of applications. It is extremely efficient at pattern recognition, for example, but not particularly good for the mathematical computation used in most of our electronic systems today. Hence, the quest for a new information processing technology will mean not only finding a new device, but also re-thinking how to apply that device and architecture to new applications and products in the future.

ACKNOWLEDGMENTS

The author gratefully acknowledges the work of the many contributors from industry, academia, NSF, NIST, and other government agencies who worked on this topic in the SIA Technical Subcommittee, the Silicon Nanoelectronics and Beyond (SNB) working groups, and the ITRS Emerging Research Devices and Emerging Research Materials sub-groups. In particular, the author would like to thank G. Bourianoff from Intel and R. Cavin and V. Zhirnov from SRC for their major contributions.

REFERENCES

Bennett, C.H. 1988. Notes on the history of reversible computation. IBM Journal of Research and Development 32(1): 16-23.

Bernevig, B.A. and S. Zhang. 2006. Toward dissipationless spin transport in semiconductors. IBM Journal of Research and Development 50(1): 141-148.

Bernstein, G.H., A. Imre, V. Metlushko, A. Orlov, L. Zhou, L. Ji. G. Csaba and W. Porod. 2005. Magnetic QCA systems. Microelectronics Journal 36(7): 619-624.

Cavin, R.K. and V.V. Zhirnov. 2004. Silicon nanoelectronics and beyond: reflections from a semiconductor industry-government workshop. Journal of Nanoparticle Research 6(2): 137-147.

Cavin, R.K., V.V. Zhirnov, G.I. Bourianoff, J.A. Hutchby, D.J.C. Herr, H.H. Hosack, W.H. Joyner and T.A. Wooldridge. 2005. A long-term view of research targets in nanoelectronics. Journal of Nanoparticle Research 7(6): 573-586.

Cavin, R.K., V.V. Zhirnov, D.J.C. Herr, A. Avila and J. Hutchby. 2006. Research directions and challenges in nanoelectronics. Journal of Nanoparticle Research 8(6): 841-858.

Dennard, R.H., F.H. Gaensslen, H.-N. Yu, V.L. Rideout, E. Bassous and A.R. LeBlanc. 1974. Design for ion-implanted MOSFET's with very small physical dimensions. IEEE Journal of Solid-State Circuits SC-9(5): 256-268.

Eerenstein, W., N.D. Mathur and J.F. Scott. 2006. Multiferroic and magnetoelectric materials. Nature 442(17 Aug): 759-765.

Geim, A.K. and K.S. Novoselov. 2007. The rise of graphene. Nature Materials 6(3): 183-191.

ITRS: International Technology Roadmap for Semiconductors. 2007. www.itrs.net/Links/2007ITRS/Home2007.htm

Khitun, A., D.E. Nikonov, M. Bao, K. Galatsis and K.L. Wang. 2007. Efficiency of spin-wave bus for information transmission. IEEE Transactions on Electron Devices 54(12): 3418-3421.

Liu, H. and P. Reinke. 2008. Formation of manganese nanostructures on the Si(100)-(2x1) surface. Surface Science 602(4): 986-992.

Min, H., G. Borghi, M. Polini and A.H. MacDonald. 2008. Pseudospin magnetism in graphene. Physical Review B 77(17 Jan): 041407-1:4

Moore, G.E. 1965. Cramming more components onto integrated circuits. Electronics 38(8).

Pearton, S.J., W.H. Heo, M. Ivill, D.P. Norton and T. Steiner. 2004. Dilute magnetic semiconducting oxides. Semiconductor Science and Technology 19: R59-R74.

Wang, L. and B. Li. 2007. Thermal logic gates: computation with phonons. Physical Review Letters 99(177208): 1-4

Welser, J.J., G.I. Bourianoff, V.V. Zhirnov and R.K. Cavin. 2008. The quest for the next information processing technology. Journal of Nanoparticle Research 10(1): 1-10.

Zhirnov, V.V., R.K. Cavin, J.A. Hutchby and G.I. Bourianoff. 2003. Limits to binary logic switch scaling – a gedanken model. Proceedings of the IEEE 91(11): 1934-1939.

Zhirnov, V.V. and R.K. Cavin. 2008. Charge of the heavy brigade. Nature Nanotechnology 3(7): 377-378.

Zutic, I., J. Fabian and S. Das Sarma. 2004. Spintronics: fundamentals and applications. Reviews of Modern Physics 76(2): 323-410.