Are we there yet? Looking beyond the end of scaling in the Nanometer Era

Sandip Tiwari, U. Avci, C. C. Liu, L. Xue, A. Kumar, S. Kim, and H. Silva
School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853

ABSTRACT

The major electronic applications of the coming decades and the technology that would make those applications possible are an important subject of discussion for industry and academia. Usefully employing gigantic scale of integration and working around the end of scaling underlie this subject, and in practice, the biggest challenge this faces is in control of power, designability, efficient interconnectivity, and reproducibility in a general purpose technology with provides a useful function. Many of the useful applications today have a mobile bent, i.e., one where power and volume and weight are critical. In this paper, we will establish the practical issues in pursuing such examples, and then discuss from group's work devices (back-plane and nano-scale), circuits (configurable and power-aware), technologies (three-dimensional), and architectures (configurable) that offer fruitful directions.

Keywords: nanoelectronics, three-dimensional integration, silicon, low power, configurable, adaptive

THE ISSUE AND THE APPLICATIONS

With the enormous impact that silicon electronics industry has in the varied endeavors of our life, it is but natural that a large amount of resources be committed to continuing the ever increasingly difficult scaling of the transistor technology [1-8], and the circuits and integration that we accomplish with it. To date, the benefits of this scaling have been enormous and the utility of mobile, or high performance, and other applications, is unquestioned. But, increasingly as we reach down from 100 nm to 10 nm, with device densities from 10^9 to 10^{11} cm^{-2}, interconnect densities from 10^{10} to 10^{12} cm^{-2}, and applications across the spectrum of digital, analog, and mixed-signal domain, a number of key issues arise related to maintaining the improvement in performance, cost, power, and design. Perhaps it is time to look at the question of what and if anything should be done at the end of the scaling.

One way of looking at this problem is to ask what the appealing and worthwhile applications are that require the continuing integration. And, we can then follow this with identifying the key issues in achieving the end. Mobile, with a large amount of data and data-stream processing, analog and mixed-signal applications, certainly lead to one set of applications that have found appeal with the consumer and hence sufficient market. Mixing of analog and digital devices has always been an issue because of the noise and cross-talk. As the densities have increased, finding ways to minimize the power has been a important problem along with the entire range of problems associated with design of devices and interconnects that need to function within the manufacturing tolerance and the need to be low cost. So, adaptive power, high density design, defect tolerance, mixed-signal, and cost are all important problems to look at if one wants to continue to derive cost advantages.

The issues of the high integration are, however, multi-faceted because they are based on interrelated questions of technology development, circuit and architecture development for performance and power, and of systems implementation, all within the criteria of affordability. In technology, these questions relate to the variety of device technologies that can now be found in silicon electronics. Multiple threshold voltage transistors as well as off-chip drivers, transistors for analog or high frequency operation, memory structures for non-volatile(NVRAM) or low power with high density (DRAM) or fast (SRAM), all require quite different practice of silicon technology adding up the costs of the processing. To usefully employ the electronics, one needs to mix many of these technologies. Almost all integrated chips with any digital processing require at the least a fast memory with logic with the microprocessor being a very common example. However, many of the newer applications – the video games and digital cameras, the mobile phones, the networking equipment for data transmission, the controllers in automotives to printers, etc., all require mixing of multitudes of these technologies, and in order to design in an acceptable time, require reuse of functional

Nanotechnology, Robert Vajtai, Xavier Aymerich, Laszlo B. Kish, Angel Rubio, Editors,
Proceedings of SPIE Vol. 5118 (2003) © 2003 SPIE · 0277-786X/03/$15.00

elements of the design. This system-on-chip approach has allowed inexpensive design, and make electronics affordable, but with technological complexity inherent in implementing different structures on planar silicon. Interconnections and devices need to scale together in order to truly benefit from the scaling of dimensions. Interconnect for short paths lead to RC delays, while for longer path lengths transmission line delays occur. The increase in interconnect density, the reduction in cross-section area and hence an increase in resistance, and limits in reduction of the capacitance, has led to increasing difficulties with reducing interconnect delays. Long paths typically have a reach of 0.15 mm/ps. For clocked designs, such as microprocessors, this means that ~100 ps latency exists in the longest interconnects of a chip. Clearly, there are a number of ways, either separately or together, that electronics can employ to chip away at these issues. Three-dimensional(3D) integration is one of these as it provides ways by which different technologies can be integrated together, ways by which it can provide higher interconnectivity through the use of vertical interconnections, ways by which it reduce signal path lengths and provides compactness, and ways by which it allows non-silicon technologies (passives, polymer-based, other inorganics: III-V, chalcogenides, etc.) and structures to be integrated.

Looking from the perspective of applications, there are multitudes of ways in which the evolution of electronics usage is likely to occur that three-dimensional integration can help with. Consider the large usage of RF in mixed-signal applications. Cross-talk, i.e., coupling of digital signal to analog elements and the need to employ a largely digital transistor design for a high frequency element are two such issues. 3D addresses these by allowing the use of ground planes to isolate electric coupling and allows freedom in design of the analog elements, such as a transistor with a metal gate to reduce gate resistance, by placing these devices on a separate device plane. In computing, embedding memory, use of large caches to reduce cache-miss penalties, use of specialized embedded DRAM processors with large data manipulation capability, reducing the longest and the average interconnect lengths, use of high interconnection grid in a symmetrically-multiprocessor with a cellular architecture, are few of the ways by which systems can be improved in performance while taking advantage of their density.

This paper will largely focus on the various facets of using the third-dimension, either actively or passively, to gain advantages that scaling has provided until now. Three-dimensional integration incorporating planar transistors connects to technology as we practice today, but provides for a path to address the issues of power, speed, cost, and systems applications. Adaptive modifications of the planar transistors offer higher scalability and functionality, higher vertical interconnectivity in between device planes can reduce interconnect delays, higher programmability using configurable elements can provide efficient signal and energy flow, higher digital-analog isolation using ground-planes can provide cross-talk improvements for mixed-signal applications, and a power-aware design can allow control of temperature and power dissipation.

Silicon Layering[9]:

Figure 1 shows a simplified diagram of the two possible ways by which one may implement silicon or silicon device layering. There are two paths to layering of wafers. In sequential processing, a silicon layer is formed on top of a processed wafer. While there are a number of processes that are being researched for this (laser recrystallization, nickel-based crystallization, epitaxial layer overgrowth), the way we implement our structures is by exfoliation. In exfoliation, a large dose of hydrogen and co-implanted species are placed inside the silicon wafer (~µm underneath the surface by implantation), and smooth surfaces of wafers attached by room temperature bonding. After strengthening of this bond, when the temperature is raised again, due to the low solubility of implanted species in silicon, the crust of silicon delaminates from the rest of the wafer and remains attached to the bonded surface. This allows, quite remarkably, an entire surface of

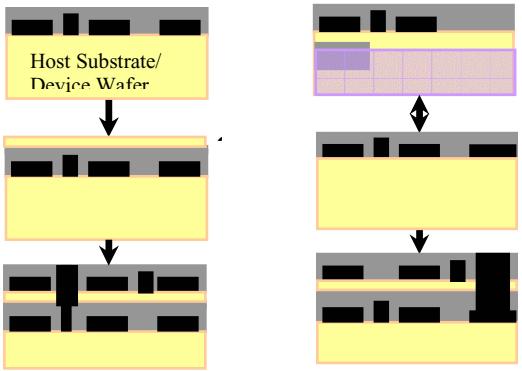
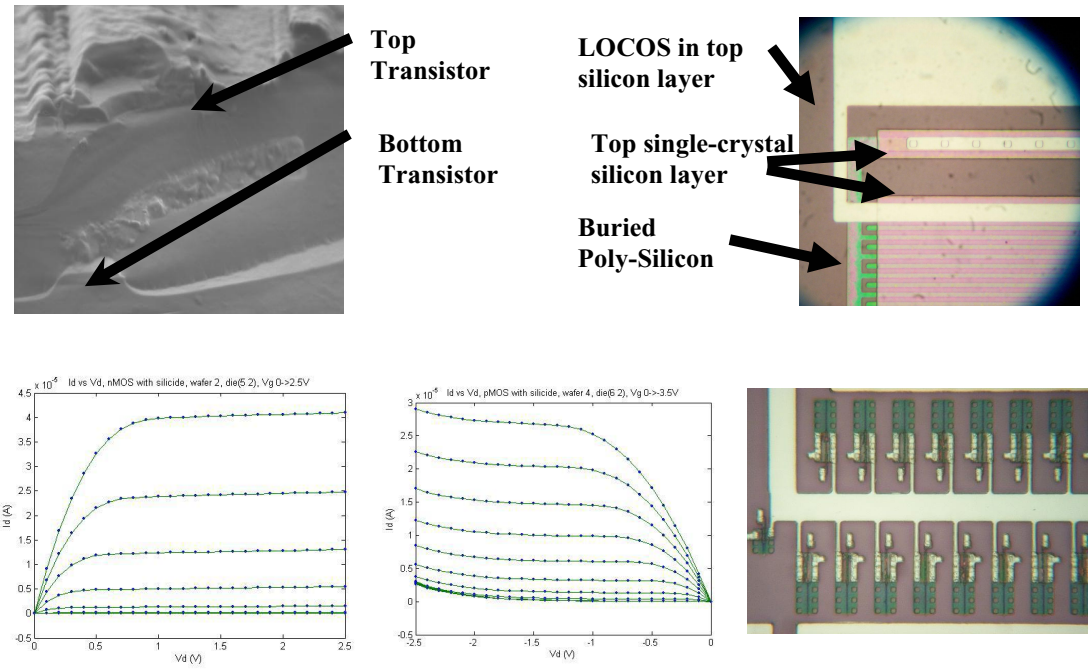


Figure 1: Sequential (left) and parallel (right) silicon layering. In the former, thin film of silicon is layered on top of a processed wafer, while in the latter processed silicon device layers are stacked.

wafer to be transferred to another wafer. The parallel processing does not allow such a process to be implemented easily because large doses of hydrogen will have unacceptable effects on reliability as well as operation of the devices. So, the preferred method in this case would be to fabricate devices in silicon-on-oxide (SOI) wafers and then use the oxide as the back-stop for the etching and layering process using a handle wafer. The first technique allows us to use our traditional lithography and processing with its similar resolution and registration. This means that the sequential processing technique allows us to achieve very high density interconnectivity in between planes at a pitch similar to that of the first level metal. The trade-off is that during the processing of the second layer, the first layer sees high temperature processing. The second process involving layering processed wafers, and hence has mechanical limitations of alignment for bonding as well as across-wafer registration effects. But, these techniques together allow us to develop and demonstrate interesting new structures.

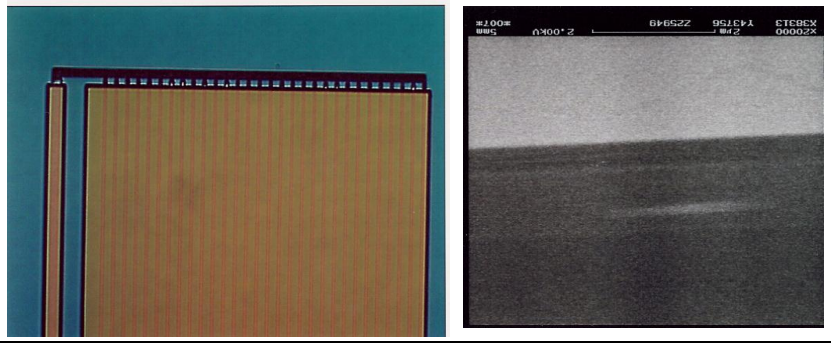
Layered CMOS[9]:
The ability to place the single-crystal silicon on top allows us to therefore continue a CMOS fabrication process. In the devices, the bottom silicon – silicon dioxide interface is a thermally grown interface, and hence the CMOS structures maintain a good interface-state density. Critical to reproducible and high-performance fabrication using this technique is the ability to obtain a reproducible silicon layer thickness. The silicon structure following exfoliation is rough (10's of nm) and needs to be chemically-mechanically polished, an essential part of reproducible bonding because of the emphasis on requiring smooth interfaces. Fig. 2 shows examples and characteristics of structures fabricated using the sequential processing described.

Figure 2: 3D planar CMOS structures fabricated using the single crystal layering technique. Picture on left is of two planes of devices. Picture on right shows example top views of layered structures.



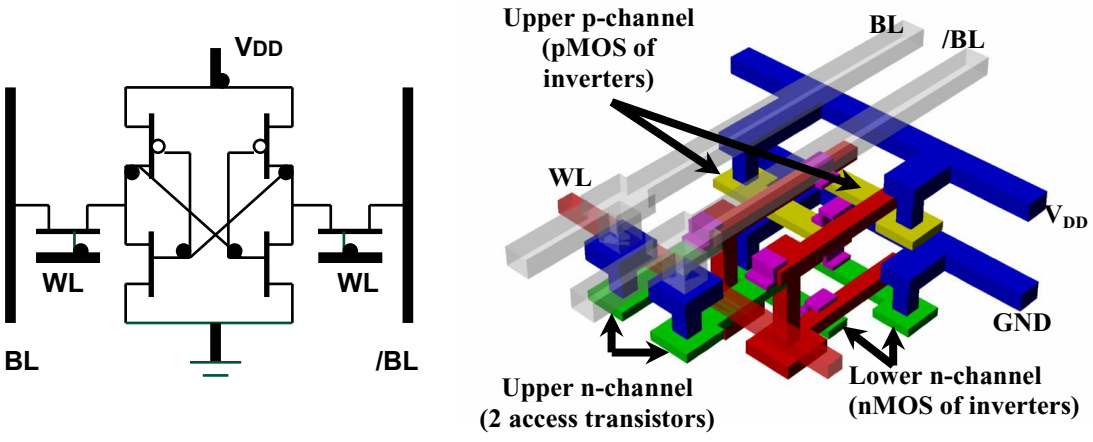
Layered CMOS with Buried Metal[10]:
Resistance of the interconnections is extremely important because of the resultant RC effects. Planar CMOS today, e.g., employs 8 levels of metal interconnects in order to achieve interconnectivity with the appropriate delay. In silicon technology, one usually does not introduce metals required for obtaining low resistances in the front-end of device processing. There is currently considerable research in metal gates in order to appropriate threshold voltage control with acceptable gate resistance at the smallest dimensions. Usually with these structures, high temperature processing is minimized and restricted to temperatures below those needed for gate oxides. We have been utilizing tungsten, deposited from tungsten hexacarbonyl, as a high temperature layering and CMOS processing compatible material in our 3D technology. While tungsten does oxidize, the W-Si-O system is highly stable, so the processing must protect the tungsten through encapsulation in order to take advantage of the stability of the ternary system. A cross-section of such buried tungsten is shown in Fig. 3.

Figure 3: Buried-tungsten 3D structures. On the left is a capacitor structure with oxidation-isolated silicon region on top. On the right is a cross-section of such layered structures.



Applications: SRAMs [11]:
Static random access memories are ubiquitous in silicon electronics because of they are fast, based as they are on a simple flip-flop or cross-connected pair of inverters, utilizing CMOS as it exists for the logic. However, SRAMs typically utilize 6 transistors (2 additional for read and write in addition to the flip-flop which consists of 4 transistors) with a fairly elaborate interconnect structure (Fig. 4). There are a total of 14 interconnections and 28 nodes/vias within the cell of the structure.

Figure 4: SRAM and one possible layout in 3D. In this layout, the mMOS inverter transistors are placed on the bottom, and the access and p-channel transistors on top.

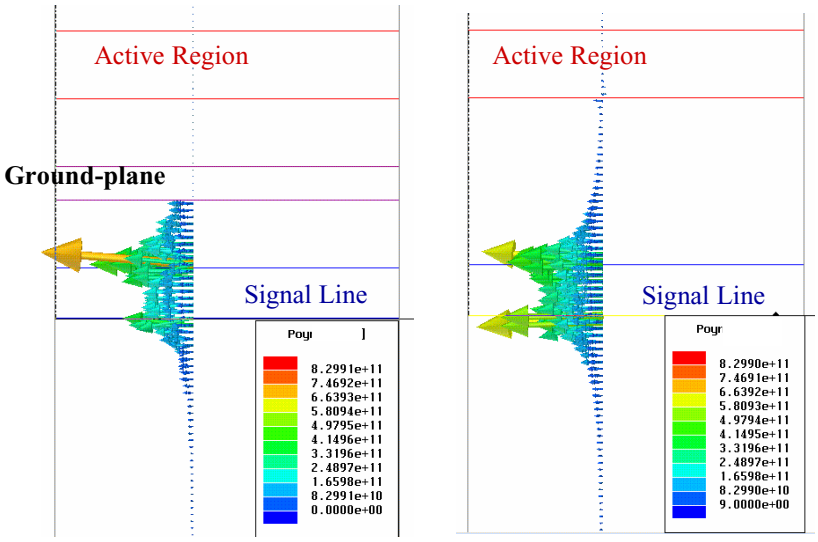


The consequence of implementing such a structure in planar CMOS technology is that a typical cell uses 80-100 squares of the lithographic feature size. And, much ingenuity in processing goes into collapsing and overlapping the interconnect structure, and cleverly eliminating design constraints in the layout of the cell. Most microprocessors, however, still utilizes half or more of the area for SRAMs that provide the on-chip cache memory hierarchy. Thus, attaining a smaller

footprint has significant area-related cost advantages. It also turns out that the smaller area and lower capacitances associated with silicon-on-insulator like technology allows one to maintain good noise margins at low power consistent with achieving high speed. Note, e.g., that in the SRAM cell shown in Fig. 4, there exist CMOS (n- and p-channel) devices simultaneously in the upper device plane, leading, at least in the form as shown in this figure, a two-plane configuration where n-channel devices of the flip-flop to be scalable for a desirable performance feature without sacrificing cell size.

Applications: Mixed-Signal[12]:
We have discussed the partitioning of functions and technology. One interesting example of this is mixed-signal applications where one may wish to design analog and digital functions in separate planes together with changes in the specific implementations of technology. Thus, analog RF structures may specifically address gate resistance issues for obtaining high maximum frequency of oscillations and have a design that provides high linearity. 3D allows one to do this, but also allows added benefits related to the coupling of digital switching noise to analog blocks. Ground planes placed to isolate coupling to the analog elements allow us to achieve this objective. An example of such isolation is shown in Fig. 5 in a simulation that shows the suppression of the energy transfer, using Poynting vector. In this simulation (implemented using HFSS), the left figure shows a ground plane isolating the active region from a signal line across which the digital signal travels. On the right are the results of the same simulations, but in the absence of the ground plane.

Figure 5: Power flow and coupling in structures with and without ground planes.



When the ground-plane is absent a substantial amount of signal energy is coupled to the active region, while the presence of a ground plane shadows and isolates the active region. The consequence of an ability to isolate such noise is that very versatile mixed-signal applications can be pursued where analog design plays as large a part as the digital design.

Applications: Adaptive and Configurable Structures[13]:
One of the many interesting attributes of three-dimensional structures is the consequence of the availability of the third dimension. An important and significant way that this occurs is the use of a buried gate and a doped substrate. By placing the gate underneath the layered silicon channel, the threshold voltage of the transistor can be modulated, thus providing controllable and adaptive properties to the transistor operation. A floating gate on the back provides a means of storing charge in the floating-gate region, thus achieving a memory using the same transistor technology. This structure also allows us to decouple the constraints of oxide required for reliability and charge retention from the scaling of the transistor for high performance operation.

Figure 6: Buried gate transistor (left), non-volatile memory (center), and configurable switch (right).

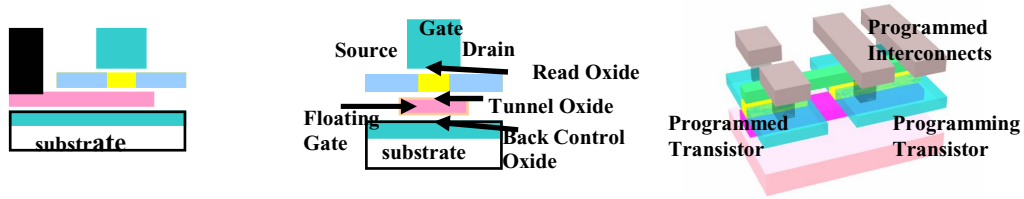
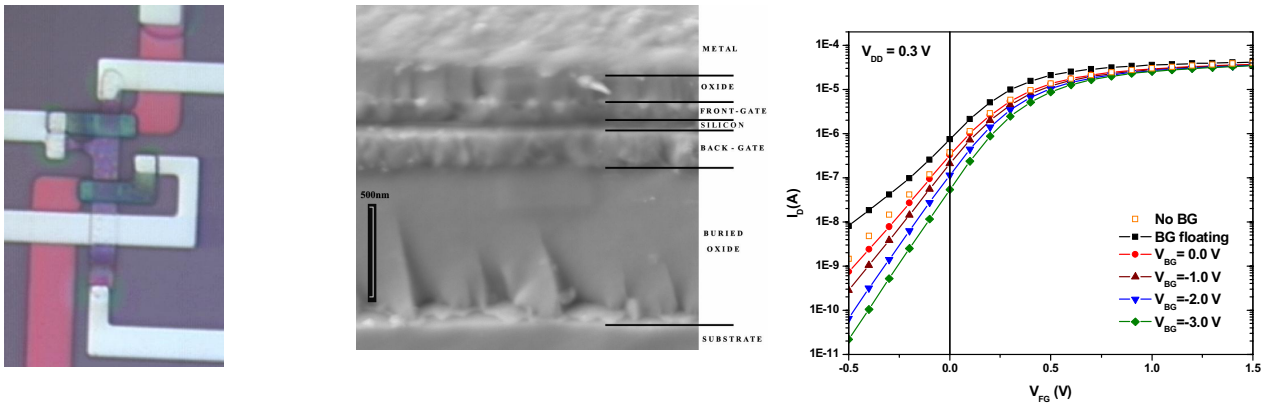


Figure 7: A back-plane circuit together with characteristics showing the control of transistor using the back-gate.

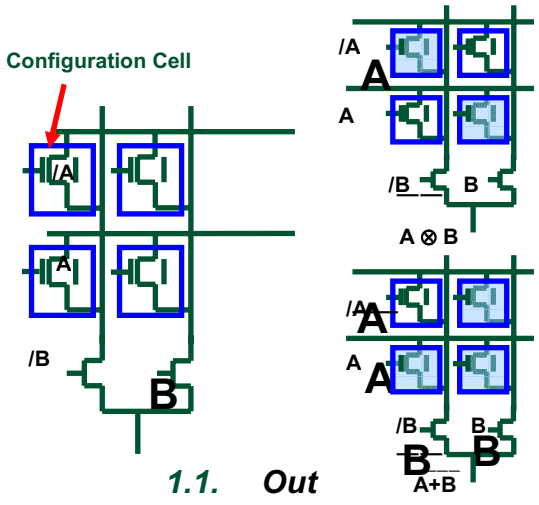


Finally, this same structure with a common floating gate between two transistors provides a method for using one transistor for programming a pass-transistor interconnection using a cross-point switch employing two programming lines. The ability to control the power and speed using the threshold voltage, achieving a non-volatile memory, and programming these properties holds great promise in adaptive power control and configurable designs.

This voltage programmability is illustrated in the experimental structure and its results. This figure shows a top view of a simple circuit structure and the characteristics of an n-channel device whose properties are controlled by the voltage applied on the back-gate. Since the devices employ thin single-crystal silicon layer, obtained by transferring them using exfoliation process, the devices are highly scalable, and the sub-threshold swing of these devices remains excellent to 10's of nm dimension. By providing for a common back gate in a circuit or a function block the excess area for providing this control is also limited.

Configurable designs that allow functional circuits and tolerance of small defect densities, and designs that allow circuits to perform multiple functions, hold a fruitful area of research and promise in addressing yield and design issues of the large densities. Fig. 8 shows an example for

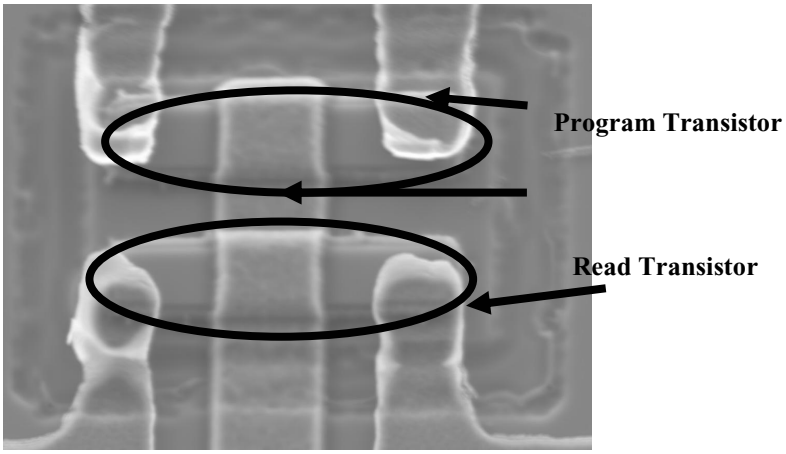
Figure 8: Configuration cell and programming using back-gate to turn off and on transistor that leads to variety of logic functions to be implemented.



such a back-gated structure implementation where a NOR and XOR functions are implemented using the same 4 transistor circuit using the programmability achieved from the back-gate. Fig. 9 shows an example of a fabricated configurable structure using the 3D approach described.

As the device count has increased over the years, and applications have moved from use of few transistors in amplifiers or logic gates, to calculators, servers, and mobile applications, the techniques used in design have shifted from careful hand-crafting to synthesis through automatic tools based on higher definition languages, optimization tools, and a smaller amount of custom manipulation. Extensive use of configurability through software is likely to be an additional step in the continuing evolution towards increasing automation with the integration.

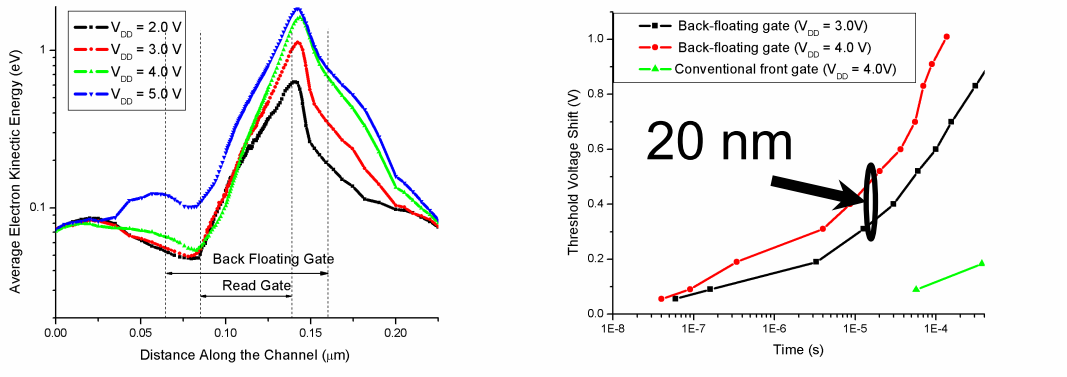
Figure 9: A scanning electron micrograph of a fabricated configurable switch.



Scaled Non-Volatile Memories:

The use of a back-gate has been mentioned as potentially providing a common technology for logic and non-volatile memory with several advantages derived from the decoupling of the scaling of the transistor control process (the effective gate control) from the constraints of charge loss. By placing the gate on the back, one achieves efficient charge injection through overlap of the floating gate region with the hot carrier region and hence an increase in speed. This allows such structures to be scalable to significantly smaller dimensions than those possible with front-floating gate structures, and yet at the same time, achieve good transistor control. The following figure shows the advantages derived for scaling and time to charge in between front-floating and back-floating structures.

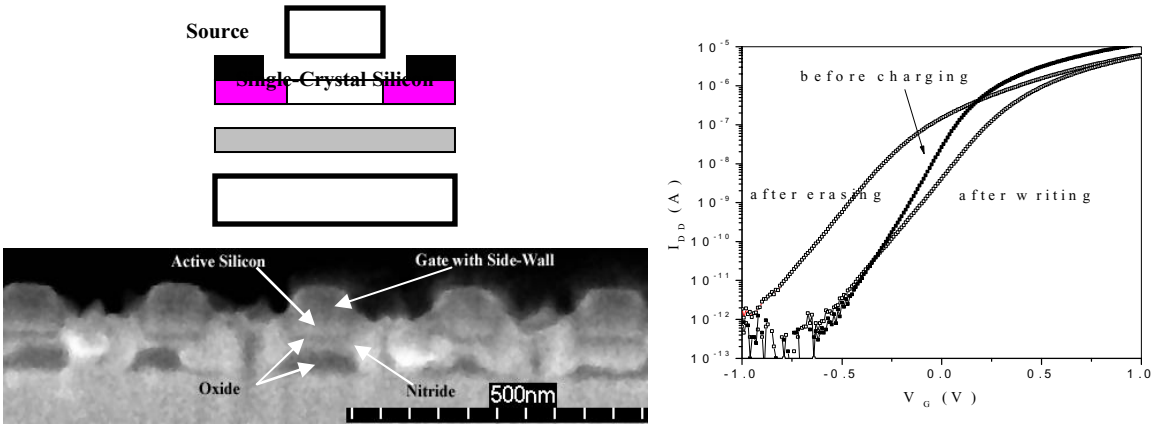
Figure 10: Hot carrier injection in the back-floating structure and the consequent differences in the scale of write times in back-gated geometries.



Use of the back-side of a channel also allows unusual possibilities in fabrication of scalable non-volatile memories in configurations that we don't currently employ. Effects due to small capacitance have important effects – such as those employed in nanocrystal memories or single electron transistors. One of the issues in use of small dimensions is the

limited statistics one works with which has consequences for the variability introduced through the variance. This is one of the severest consequences of nanodimensions in electronics. In non-volatile memories, there is an unusual opportunity for the use of defects at smallest dimensions that employs this back side principle [14,15].

Figure 11: A back defect layer memory which also operates as a transistor.



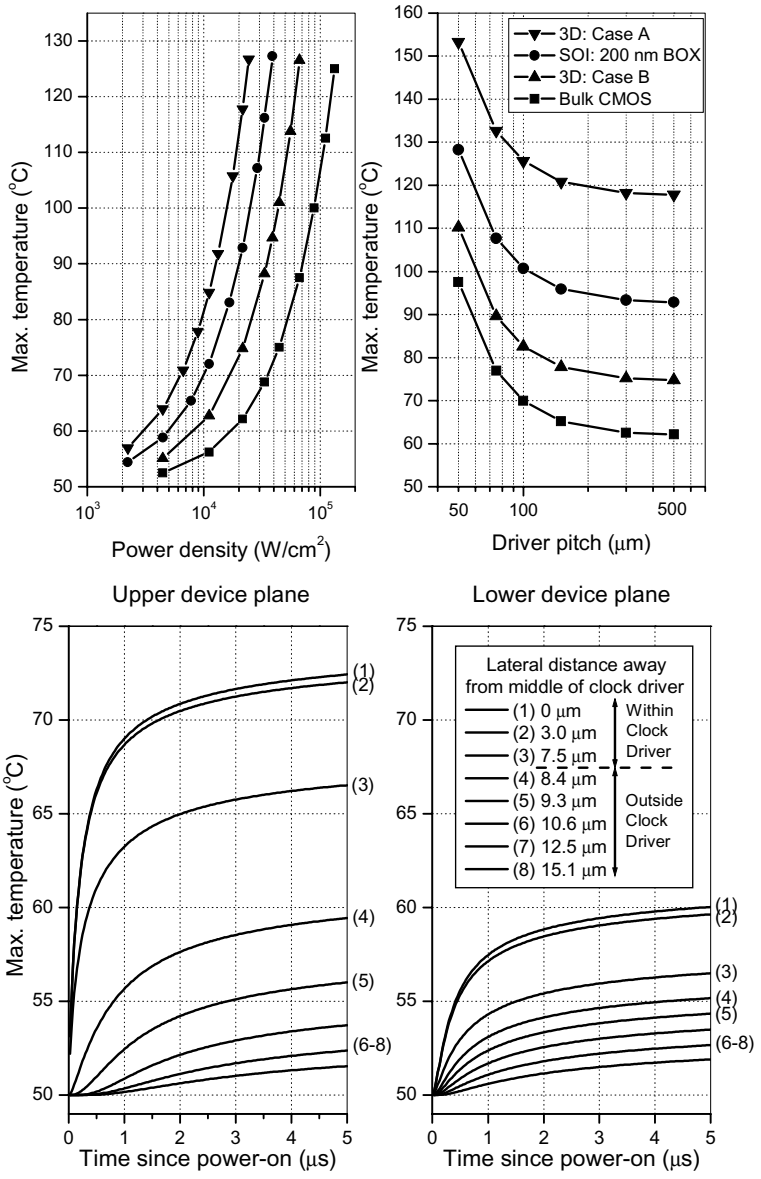
The charge is stored on defects on the back of the device structure, and the charge can be injected efficiently only at high voltages. So, these devices operate as transistors at low voltages and as memories at high voltages, and can be architecturally employed to perform both functions.

Issues with 3D Integration[16]:
That 3D holds much promise is quite clear from the snap-shot provided above. It is quite clearly one possible path to getting beyond the limits set by interconnects and the size limits of field-effect. However, like all research directions, 3D also raises interesting problems that need to be taken care of by design or technology. Besides the development of reproducible technology, two prominent ones are those related to power and yield. These are problems for planar technology too (and silicon-alternatives), but become accentuated in 3D because of limited thermal conductivity of insulators (SiO_2 is nearly factor 100 worse than Si), and the yield limitations of a new technology. It is projected that by the end of this decade, manufacturability would be sufficiently advanced to make 14 cm^2 as the ideal cost-effective size of a chip. Our lithography tools and lateral connectivity delays typically constrain us to work within a foot-print which is at most 2.5 cm on the side, except in specialized applications such as CCDs, etc., where steppers strap the interconnects at the edge of each chip. If these predictions are really true, they seem to indicate that at least 2 or 3 layer 3D circuits should be feasible with reasonable yield within the 2.5 cm x 2.5 cm footprint.

The thermal issue, however, is more bothersome for all electronic designs. In particular, the acuity of this problem arises from synchronous design style that is mainstay of technology today. Clocks, usually distributed across the chip, consume nearly 70 percent of the power so that the slew is small and the entire chip operates closely in unison. Clock drivers are individual transistor elements with large drive capability, and therefore are the highest power density elements and show the highest temperatures in a chip. Use of SOI technology has led to an increase in these temperatures because the oxide underneath conducts heat poorly. Figure 12 shows a calculation of the temperature rise and time transient effects in clock drivers which is representative of silicon technology at 180 nm dimensions.

These figures show that heat propagation causes unusual effects. Temperatures can be significantly higher because of conductivity effects, and time delays occur because of slow propagation. However, there are interesting advantages also visible in these calculations. Interconnectivity in between planes places elements of high thermal conductivity between the planes. These have the attribute of mitigating the effect of the oxide. Thus, small interconnect lengths, e.g., the case B of Figure 12, actually improves on the SOI case because the small interconnect lengths behave as heat pipes for extracting the heat away from the top plane. The bottom figure shows that the upper device plane with heat propagating laterally due to the oxide heat barrier underneath, shows delay effect in the thermal transient. Thus, one implication of this behavior is that one would have to carefully look at the heat dissipation properties as one partitions the functional and technology design of the 3D implementation.

Figure 12: Clock driver related behavior in 3D. Figure on top shows temperature as a function of power density and driver pitch effect in a variety of silicon implementations. The figure on bottom shows the time-transient effect due to heat propagation in 3D structures.



Other Comments:

We had discussed the two possible pathways to 3D integration – either as sequential with the dense interconnectivity or parallel with lower interconnectivity. Lower interconnectivity clearly has implication from the heat perspective, but it also allows a bus-oriented design such as for system-on-chip applications. The biggest care one has to exercise in such applications is that heat dissipation is controlled and that one designs to take a large advantage of the high interconnectivity bus. Embedded DRAM, or SRAM, integrated to logic, or mixed-signal design combining analog and digital planes, are two examples of such application. The sequential, on the other hand provides wider density-oriented freedom, but at the cost of the higher temperature effects of processing.

Summary:

This paper provided a rather diverse view of what 3D technology entails in its promise for devices and circuits and what the limitations of the technology appear to be at this moment in time and its promise as we reach the end of scaling. The approach to silicon's dimensional limits does not mean that we have reached the end of what we can accomplish with silicon. Higher integrations, low interconnect penalty, new design styles, and high volume and area efficiency that 3D integration will provide is likely to lead to continuing evolution in silicon applications.

Acknowledgements:
This effort in three-dimensional integration has benefited tremendously from the support provided by DARPA (N66001-00-C-8003) through the AME and HGI program and Cornell-IBM (D. J. Frank, K. Guarini, and R. McFeely) collaboration. Support from NSF and SRC has been instrumental in exploring anoscale structures for devices, new applications and circuits using the devices, and new 3D approaches. Cornell Nanofabrication Facility was vital for the fabrication effort summarized here.

REFERENCES:

[1] R. D. Isaac, "The future of CMOS Technology," *IBM J. Res. Develop*., 44(3), 369-378, 2000
[2] J. D. Meindl, Q. Chen, and J. A. Davis, "Limits on Silicon Nanoelectronics for Terascale Integration," *Science*, 293, 2044-2049, 2001
[3] D. J. Frank, R. H. Dennard, E. Nowak, P. M. Solomon, Y. Taur, and H-S P. Wong, "Device Scaling Limits of Si MOSFETs and their Application Dependencies," *Proc. of IEEE*, 89(3), 259-288, 2001
[4] J. C. Ellenbogen and J. C. Love, "Architectures for molecular electronic computers. I. Logic structures and an adder designed from molecular electronic diodes," *Proc. of the IEEE* , 88(3), 386-426, 2000
[5] S. Carroll, "A Complete Programming Environment for DNA Computation," *Workshop on non-Silicon Computation*, NSC-1, Cambridge, 2002
[6] S. A. Wolf, D. D. Awschalom, R. A. Buhrman, J. M. Daughton, S. von Molnar, M. L. Roukes, A Y. Chtchelkanova, and D. M. Treger, "Spintronics: A Spin-Based Electronics Vision for the Future," *Science* 294(16): 1488-1495, 2001
[7] A. Bachtold, P. Hadley, T. Nakanishi, and C. Dekker, "Logic Circuits with Carbon Nanotube Transistors," *Science* 294, 1317-1320, 2001
[8] S. Tiwari, "Silicon Electronics at the Nanoscale: Devices, Circuits and Architecture," *Tech. Dig. of 7'th Int. Conf. on Nanometer-scale Science & Technology & 21'st European Conf. on Surface Science* June 2002
[9] L. Xue and S. Tiwari, "Multi-Layers with Buried Structures (MLBS): An Approach to Three-Dimensional Integration," *Tech. Dig. of IEEE International SOI Conference*, 117 Oct. ,2001
[10] Hong Seung Kim, L. Xue, A. Kumar and S. Tiwari "Fabrication and Electrical Properties of Buried Tungsten Structure for Direct Three Dimensional Integration," *Solid State Devices Meeting*, Japan, 2002
[11] C. C. Liu, and S. Tiwari, "Application of 3D CMOS Technology to SRAMs," IEEE *Int'l. SOI Conf*., Charlottesville (2002)
[12] L. Xue, C. C. Liu, H S Kim, S (Kevin) Kim, and S Tiwari, "Three-Dimensional Integration: Technology, Use, and Issues for Mixed-Signal Applications," *IEEE Trans. on Electron Devices*, Mar. 2003
[13] A. Kumar and S. Tiwari, *Tech. Dig. of IEEE Silicon Nanoelectronics Workshop*, June 2002
[14] H. Silva and S. Tiwari,, "A Novel Silicon Based Transistor-Memory Device, *APS March Meeting*," N8.009, Mar. 5, 2003
[15] H. Silva and S. Tiwari, "A Scalable Nano-Transistor and Memory using Back-Side Trapping," *Tech. Dig. Of IEEE Silicon Nanoelectronics Workshop*, June (2003)
[16] C. Liu, J. Zhang, A. K. Datta, and S. Tiwari, "Heating Effects of Clock Drivers in Bulk, SOI, and 3D CMOS," *IEEE El. Dev. Letters*, Dec. 2002