Future Directions

Taking Silicon to the Limit: Challenges and Opportunities

by Tsu-Jae King

Silicon-based CMOS transistors can be scaled well into the sub-10 nm regime. However, new materials and processes, in conjunction with advanced transistor structures, will be needed for nanometer-scale MOSFETs to meet performance specifications in the International Technology Roadmap for Semiconductors (ITRS). This paper discusses challenges for achieving target performance metrics at the end of the Roadmap, and approaches to overcoming them.

The steady miniaturization of the metal-oxide-semiconductor field-effect transistor (MOSFET) with each new generation of complementary-MOS (CMOS) technology has yielded continual improvements in integrated-circuit performance (speed) and cost per function over the past several decades, to usher in the Information Age. Continued transistor scaling will not be as straightforward in the future as it has been in the past, however, because fundamental materials and process limits are rapidly being approached.[1] New materials and processes, as well as non-classical transistor structures, will be necessary in order to extend CMOS technology to the last node of the ITRS.[2] Minimization of leakage current, parasitic resistance, and capacitance to minimize power consumption and maximize circuit performance, and reduction in device-to-device variability to increase yield (and thereby lower cost), will be key challenges for sustaining the rapid growth of the industry to usher in the age of ambient intelligence and ubiquitous computing. This paper discusses recent CMOS technology developments and remaining work needed to address these challenges.

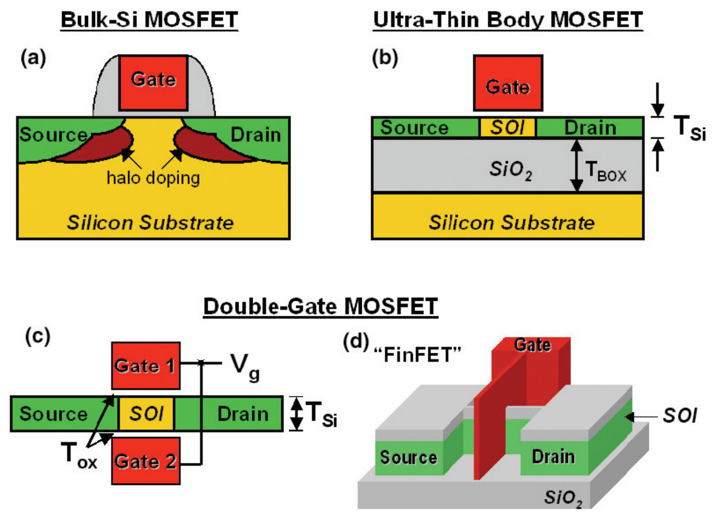


FIG. 1. Schematic diagrams of MOSFET structures: (a) classical bulk-Si, (b) ultrathin-body (UTB), (c) double-gate (DG), and (d) FinFET.

Advanced Transistor Structures and Materials

In order to scale the classical bulk-Si MOSFET structure (Fig. 1a) down to the 10 nm physical gate length (L_g) regime, heavy halo and channel doping (greater than 1×10^{18} cm^{-3}) will be required to suppress leakage current and short-channel effects.[3] As a result, field-effect carrier mobilities will be degraded, resulting in incommensurate improvements in transistor drive current with L_g scaling.[4] Thin-body transistor structures (Figs. 1b and 1c)[5] rely not on heavy channel doping but on a sufficiently thin body/channel region ($T_{Si} < L_g$) to limit leakage current. The use of a lightly doped or undoped channel provides immunity to variations in threshold voltage (V_T) resulting from statistical dopant fluctuations in the channel, as well as enhanced carrier mobility for higher transistor drive current because of the lower transverse electric field in the inversion layer.[4] Therefore, thin-body MOSFETs offer improved circuit performance as compared to the bulk-Si transistor structure (Fig. 2).[6] To provide a means for adjusting V_T without channel doping during the manufacturing process, a tunable-work-function gate technology is necessary. For thin-body CMOSFETs, the required range of gate work function (Φ_M) tunability is from 4.5 eV to 5.0 eV.[7]

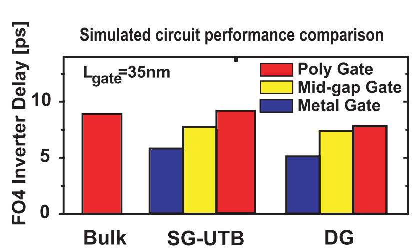


FIG. 2. Loaded-inverter delay comparison of bulk-Si vs. UTB vs. DG CMOS technologies, obtained through mixed-mode simulation using realistic device structures based on ITRS specifications.[6]

Advanced Transistor Structures

Ultrathin-body (UTB) FET—The body thickness T_{Si} must be less than ~$L_g/3$ in a UTB FET in order to adequately suppress leakage current.[5] Because of quantum confinement effects, V_T becomes a sensitive function of T_{Si} for thicknesses below 5 nm.[8] Also, carrier mobilities are degraded due to enhanced interface roughness scattering for T_{Si} < 4 nm.[9] For these reasons, it may be difficult to scale the UTB FET structure to below 12 nm L_g, unless techniques for achieving uniformly thin films with atomically smooth surfaces/interfaces are used. One example is the "Silicon-on-Nothing" fabrication process,[10] in which the ultrathin Si channel and the buried oxide are defined by epitaxy on a bulk-Si substrate, so that thickness control can be as fine as a single atomic layer.

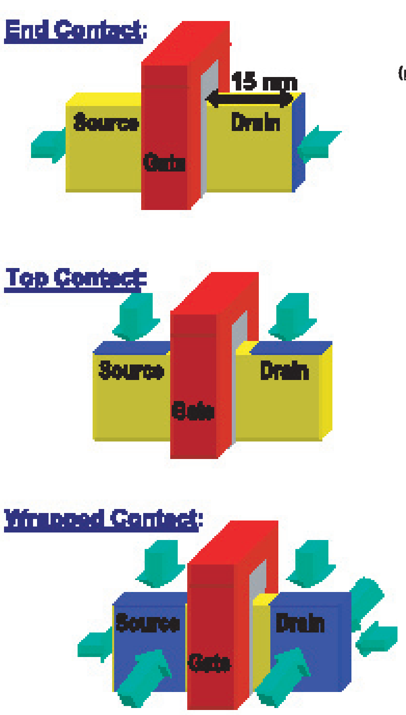
Double-gate (DG) FinFET—The quasi-planar FinFET (Fig. 1d) offers the superior scalability of the DG MOSFET structure together with a process flow and layout similar to that of the conventional MOSFET.[11] Hence, it has been investigated by many companies.[12-14] FinFETs with gate lengths down to 10 nm have already been demonstrated and exhibit excellent control of short-channel effects.[14,15] One advantage of this vertical transistor structure is that it is relatively immune to gate line-edge roughness, a major source of variability in planar nanoscale FETs.[16] FinFET performance variability due to variations in fin width is a potential issue, but can be minimized by using a spacer lithography process.[17] In order to optimize the trade-off between parasitic series resistance and parasitic gate capacitance, a gate-underlapped structure (in which the electrical channel length is larger than the physical gate length) will be required to achieve peak circuit performance for sub-20 nm L_g.[18] Parasitic series resistance and contact resistance will therefore ultimately limit FinFET performance in the nanoscale regime (Fig. 3).[19] The use of thick source/drain (S/D) regions, *e.g.* formed by selective growth of Si[12], $Si_{1-x}Ge_x$, or Ge,[20] can help to alleviate this issue, particularly if low specific contact resistivity ($\rho_c < 10^{-8}$ Ω-cm^2) contacts can be formed by silicidation/germanidation of the S/D fin surfaces. It should be noted that the width of the gate-sidewall spacers, which isolate the gate electrode from the raised-S/D regions, must be carefully optimized for peak circuit performance.[21]

Advanced transistor structures can be applied to improve the scalability of memory devices as well, to achieve very high density, non-volatile information storage. FinFET SONOS (silicon-oxide-nitride-oxide-silicon) memory devices have already been demonstrated, and exhibit excellent retention and endurance characteristics.[22] Because the effective gate-dielectric thickness of the ONO stack is relatively large (~10 nm), the body thickness of a SONOS memory device must be even thinner than for a logic device (a significant challenge for fabrication), in order to adequately suppress leakage current (Fig. 4).[22] If the two gate electrodes are electrically isolated (*e.g.*, by applying a chemical mechanical polishing step or a masked etch step during the fabrication process), then 2-bit storage is possible, to further increase storage density.[23] In order to allow each bit to be distinguished, asymmetric gates are required (n^+ poly-Si for the front gate, and p^+ poly-Si for the back gate). These can be obtained in a straightforward manner, via high-tilt-angle implants to dope the front and back gates separately.[24]

Back-Gated (BG) UTB FET—Power consumption will be a primary design constraint for sub-65 nm CMOS technologies, so that active leakage (V_T) control will be necessary for optimization of energy *vs.* delay trade-offs in future ULSI systems. The V_T of a FinFET cannot be dynamically changed; however, if the two gate electrodes are electrically isolated so as to allow independent operation,[25] the FinFET can be operated as a back-gated UTB FET with the capability for dynamic V_T control. It should be noted that the optimal BG UTB FET design employs significantly different gate-oxide thicknesses for the front and back gates.[26] Techniques such as selective (tilted) implantation of nitrogen,[27] oxygen,[28] or argon[29] to simultaneously grow gate oxides of different thicknesses can be used for the FinFET structure. A planar BG UTB FET structure may ultimately be more area-efficient because the back gate would be routed in a separate layer than the front gate. The development of a cost-effective fabrication process for the planar BG UTB FET with self-aligned gates (for optimal performance), as well as the need to achieve $T_{Si} < 4$ nm with excellent uniformity and oxide-interface quality, remains a challenge.

Extending the Roadmap—The scaling limit of the Si MOSFET is well below 10 nm L_g (depending on the leakage current specification and power-supply voltage), and corresponds to the point where direct tunneling of carriers from the source to the drain in the off state becomes prohibitively large.[30] Practically, the MOSFET scaling limit will be determined by the degree to which the channel film thickness can be controlled in a manufacturing process. Nanofabrication techniques (*e.g.* self-assembly) may be useful for achieving the uniformly thin and smooth channel films required to reduce performance variations to an acceptable level, as well as for improving critical-dimension (*i.e.*, L_g) control. Further scaling of L_g can also be enabled by improving the transistor design (*e.g.* so that the effective channel length or V_T is larger in the off state than in the on state) and/or by employing an alternative semiconductor material (one with lower dielectric permittivity to reduce drain-induced barrier lowering). Clearly, opportunities abound for innovations by technologists and device designers alike to extend transistor scaling toward atomic dimensions.

(continued on next page)



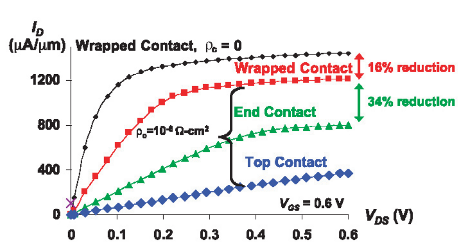


Fig. 3. Impact of S/D contact structure on FinFET drive current, obtained by 3-D device simulation.[19] (L_g = 18 nm, T_{Si} = 10 nm, T_{ox} = 5Å, S/D profiles optimized)

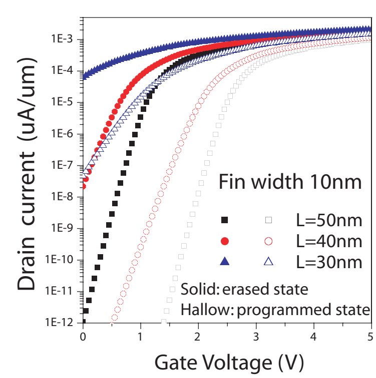


Fig. 4. Transfer characteristics for FinFET SONOS memory (ONO = 3 nm/6.1 nm/4.8 nm), from 2D device simulation.[22] An adequately large difference in currents for erased vs. programmed states is achieved for L_g < 40 nm if a sufficiently thin (10 nm) body is used.

King
(continued from previous page)

Tunable-Work-Function Gate Materials

Two metal-gate technologies have been demonstrated to provide a range of Φ_M tunability suitable for application to future thin-body CMOSFETs: molybdenum (Mo) and fully-silicided NiSi. Mo is compatible with a conventional ("gate first") fabrication process;[31] it can be deposited by sputtering or chemical vapor deposition, patterned by conventional reactive ion etching, and is thermodynamically stable on SiO_2. Nitrogen (N) implantation followed by thermal annealing to form a Mo nitride layer at the gate-oxide interface has been shown to be an effective way to controllably lower the effective Φ_M of a Mo gate electrode on SiO_2 gate dielectric, from 5 eV down to 4.4 eV (Fig. 5).[32] Care must be taken avoid Mo^+ diffusion[33] and gate-oxide damage during sputter deposition;[34] the Mo gate film thickness and N implant energy must be carefully co-optimized to minimize damage (due to implant straggle) to the underlying gate dielectric.[35] A capping layer (*e.g.,* of TiN) deposited *in situ* is beneficial to prevent out-diffusion of the implanted N during thermal annealing, to maximize the Φ_M shift for a given implant dose and to improve uniformity,[36] which is critical for precise V_T control.

The work function of a fully silicided (FUSI) gate material can be adjusted via doping of the precursor Si gate material.[37] Researchers have recently demonstrated that Φ_M for a FUSI NiSi gate on SiO_2 gate dielectric can be tuned over a significant range[38] (from 4.5 eV to 4.9 eV, for dopant implant doses up to ~3 × 10^{15} cm^{-2} [39]) and have successfully applied this gate technology to fabricate CMOS FinFETs with nearly symmetrical V_T's.[40] The NiSi is formed at low temperature ≤500°C) and cannot withstand high annealing temperature; hence, the silicidation must be the last thermal processing step in the transistor fabrication process. The intrinsic tensile stress (~0.8 GPa) in a NiSi gate induces tensile strain in the Si channel for narrow-width UTB FETs, which enhances both electron and hole carrier mobilities and hence drive current.[41]

It should be noted that Φ_M for a metal gate electrode can vary significantly with the gate-dielectric material.[42] In addition, process integration challenges will change with the gate-dielectric material: Ni atoms in a FUSI NiSi gate can easily penetrate HfO_2 during the silicidation process, leading to yield and reliability problems;[43] N implanted into a Mo gate can easily penetrate HfO_2 during subsequent annealing steps and degrade the oxide-silicon interface and hence transistor performance.[44] Therefore, a metal-gate technology must be tailored specifically to the gate-dielectric material. For HfO_2, FUSI HfSi is a promising gate material (with Φ_M tunable in the range 4.23 eV to 4.87 eV) that is stable at high temperatures and, therefore, compatible with a conventional ("gate-first") planar CMOS fabrication process.[45]

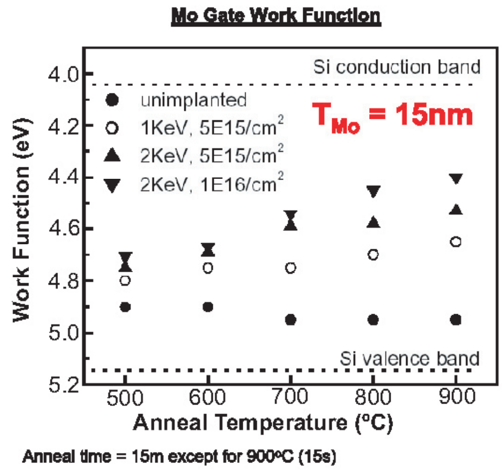


FIG. 5. Effect of N^+ implant and annealing conditions on the effective Φ_M of a Mo gate on SiO_2.[31]

Performance Enhancement Approaches

Alternatives approaches to transistor scaling for continued improvements in system performance and reductions in cost and power consumed per function will ultimately be needed to sustain the rapid growth of the semiconductor industry through the first half of this century. Process technology innovations which improve transistor drive current (I_{on}) without sacrificing off-state leakage (I_{off}) can further improve the performance vs. power trade-off. Device innovations can provide circuit designers with better "building blocks" to enable more efficient designs. Examples of these are discussed below.

Carrier-mobility enhancement—Techniques for increasing the average velocity of carriers in the channel—without significantly impacting cost and device reliability—will be necessary in order for the industry to maintain its historic 17%-per-year performance improvement rate.[2] Approaches to enhancing carrier mobility include the use of a strained capping layer,[46] a strained gate electrode,[47] or strained S/D regions (using epitaxial $Si_{1-x}Ge_x$ [48] or silicide [49]), and optimization of the channel surface crystal orientation and current flow direction;[50] indeed, some of these methods are already used in state-of-the-art CMOS products today. In the future, these techniques must be adapted to advanced transistor structures, with manageable device-parameter dependencies (*i.e.,* variation with transistor channel length and width, and with S/D length) for effective circuit design.

Metallic-source/drain technology—In the sub-10 nm L_g regime, intrinsic variation due to random discrete dopants in the S/D regions will cause large variations in I_{on} and I_{off} .[16] The use of metallic-S/D regions rather than doped-S/D regions can eliminate this issue. Sub-20 nm L_g CMOSFETs with silicide S/D have been successfully fabricated using the UTB structure to achieve low leakage current.[51] The primary challenge for metallic-S/D technology is achieving sufficiently low (≤0.1 eV [52]) Schottky barrier height Φ_b to meet the Roadmap I_{on} specifications. Techniques proposed for lowering Φ_b include passivation of extrinsic interface states,[53] straining the Si channel,[54] and using very heavily doped "tips" formed by silicidation-induced impurity segregation[55] (which is susceptible to random discrete dopant effects).

Negative differential resistance devices—Negative differential resistance (NDR) devices are a prime example of alternative semiconductor devices that can potentially be used to reduce the power consumption and cost of integrated circuits (ICs). The defining feature of an NDR device is that the current flowing between two of its terminals actually decreases as the voltage difference between those two terminals increases over a range of voltages. A key figure of merit for NDR devices is the "peak-to-valley" current ratio (PVCR). The higher the PVCR, the better: A high "peak" current is needed for fast and reliable circuit operation, while a low "valley" current is needed to minimize power consumption. Silicon-based NDR devices typically exhibit a PVCR no greater than 10 at room temperature,[56] decreas-

(continued on page 42)

Future Directions

King

(continued from page 40)

ing significantly as the temperature is increased.

IC manufacturers have actively investigated devices that exhibit significant NDR behavior since the invention of the Esaki diode.[57] This is because such devices used together with conventional transistors result in more efficient circuits because fewer elements are needed to implement a function.[58] Many innovative circuit designs and applications have been proposed in the literature (see Ref. 58 for an overview) to take advantage of NDR devices, including: compact static memory (SRAM),[59] self-latching logic, analog-to-digital conversion, shift registers, oscillator elements, and multi-valued logic. To date, technological obstacles have hindered their widespread use in silicon-based ICs, however. This is because high-performance NDR devices typically require highly specialized (*i.e.*, expensive) fabrication processes and/or exotic materials, so that they cannot be easily integrated with conventional CMOS devices. Thus, the development of a high-performance (high-PVCR), CMOS-compatible NDR device would constitute a breakthrough advancement in IC technology and have a significant impact on the industry.

Summary

As compared to the classical bulk-Si MOSFET structure, thin-body transistor structures achieve a better trade-off between performance and power consumption, and also can provide immunity to random variations associated with the discreteness of dopant atoms (if a tunable-Φ_M gate material and metallic S/D are used) and line-edge roughness effects. Therefore, they will likely be used to scale L_g to below 10 nm. Techniques for achieving uniformly ultrathin (<5 nm) channel films with atomically smooth surfaces will be needed to extend transistor scaling further, toward 1 nm L_g. Precise control of interfacial properties (not only at Si-dielectric interfaces, but also at metal-dielectric interfaces and Si-metal interfaces) will be critical for achieving high performance with good uniformity in nanometer-scale transistors. The need has never been greater for innovations in process technology, materials, and device design to sustain the Si revolution. ■

Acknowledgments

I have learned much from my colleagues at UC Berkeley, especially Chenming Hu, Jeffrey Bokor, Bora Nikoli□, Hideki Takeuchi, Leland Chang, Yang-Kyu Choi, Jakub Kedzierski, Pushkar Ranade, Min She, Sriram Balasubramanian, Daewon Ha, Hei Kam, Kyoungsub Shin, and Hiu Yung Wong. Research funding from DARPA (Advanced Microelectronics Program), SRC (Advanced Devices and Technology Program), the MARCO Focus Center for Advanced Materials, Structures, and Devices, and the MARCO Focus Center for Circuits, Systems, and Software is gratefully acknowledged.

References

1. D. J. Frank, *et al., Proc. IEEE*, **89**, 259 (2001).
2. Semiconductor Industry Association, *International Technology Roadmap for Semiconductors*, 2003 Edition, http://public.itrs.net/Files/2003ITRS/Home2003.htm.
3. A. Hokazono, *et al., Tech. Dig. - Int. Electron Devices Meet.*, p. 639 (2002).
4. D. Antoniadis, *Technical Digest-Very Large Scale Integration (VSLI) Symposium, IEEE, p. 2 (2002).*
5. Y.-K. Choi, *et al., IEEE Electron Device Lett.*, **21**, 254 (2000).
6. S. Tang, *et al., Tech. Dig. - IEEE Int. Solid-State Circuits Conf.*, p. 118 (2000).
7. L. Chang, *et al., Tech. Dig - Int. Electron Devices Meet.*, p. 719 (2000).
8. Y.-K. Choi, *et al., IEEE 58th Device Research Conf.*, p. 85 (2001).
9. K. Uchida, *et al., Tech. Dig. – Int. Electron Device Meet.*, p. 47 (2002).
10. M. Jurczak, *IEEE Trans. Electron Devices*, **47**, 2179 (2000).
11. Lindert, *et al., IEEE Electron Device Lett.*, **22**, 487 (2001).
12. J. Kedzierski, *et al., Tech. Dig. – Int. Electron Devices Meet.*, p. 437 (2001).
13. F.-L. Yang, *et al., Tech. Dig. – VLSI Symp.*, p. 104 (2002).
14. B. Yu, *et al., Tech. Dig - Int. Electron Devices Meet.*, p. 251 (2002).
15. F.-L. Yang, *et al., Tech. Dig. – VLSI Sympos.*, p. 196 (2004).
16. A. R. Brown, *et al., IEEE Trans. Nanotechnology*, **1**, 195 (2002).
17. Y.-K. Choi, *et al., IEEE Trans. Electron Devices*, **49**, 436 (2002).
18. S. Balasubramanian, *et al., Proc. Silicon Nanoelectronics Workshop*, p. 16 (2003).
19. H. Kam, *et al., Proc. Silicon Nanoelectronics Workshop*, p. 9 (2004).
20. N. Lindert, *et al., IEEE Int. SOI Conf.*, p. 111 (2001).
21. S. Zimin, *et al., Proc. 5th Int. Conf. Solid-State and Integrated Circuit Technology*, p. 188 (1998).
22. P. Xuan, *et al., Tech. Dig. – Int. Electron Device Meet.*, p. 609 (2003).
23. K. H. Yuen, *et al., IEEE Electron Device Lett.*, **24**, 518 (2003).
24. L. Mathew, *et al., IEEE Int. SOI Conf.*, p. 109 (2003).
25. L. Mathew, *et al., IEEE Int. Conf. on Integrated Circuit Design and Technology*, p. 97 (2004).
26. S. Balasubramanian, *et al., IEEE Int. SOI Conf.*, p. 27 (2004).
27. B. Doyle, *IEEE Electron Device Lett.*, **16**, 301 (1995).
28. Y.-C. King, *et al., Tech. Dig. – Int. Electron Device Meet.*, p. 585 (1998).
29. M. Togo, *et al., Tech. Dig. – Int. Electron Device Meet.*, p. 347 (1998).
30. J. Wang, *et al., Tech. Dig. – Int. Electron Device Meet.*, p. 707 (2002).
31. D. M. Brown, *et al., Solid-State Electron.*, **11**, 1105 (1968).
32. P. Ranade, *et al., Tech. Dig. – Int. Electron Device Meet.*, p. 363 (2002).
33. J. D. McBrayer, R. M. Swanson, and T. W. Sigmon,*, J. Electrochem. Soc.*, **133**, 1242 (1986).
34. T. Amazawa, *et al., J. Electrochem. Soc.*, **145**, 1297 (1998).
35. T.-J. King, *Semiconductor Interface Specialists Conf.*, p. S 2.1 (2002).
36. K. Shin, *et al.*, in preparation.
37. M. Kakumu, *et al., Tech. Dig. – VLSI Symp.*, p. 30 (1984).
38. M. Qin, *et al., J. Electrochem. Soc.*, **148**, G271 (2001).
39. J. Kedzierski, *et al., Tech. Dig. – Int. Electron Device Meet.*, p. 13.3.1 (2003).
40. J. Kedzierski, *et al., Tech. Dig. – Int. Electron Device Meet.*, p. 247 (2002).
41. Z. Krivokapic, *et al., Tech. Dig. – Int. Electron Device Meet.*, p. 445 (2003).
42. Y. Yeo, *et al., Tech. Dig. – VLSI Symp.*, p. 49 (2001).
43. J. Schaeffer, *et al.*, Paper D4.1, presented at the 2004 MRS Spring Meeting, San Francisco, CA, April 12-16, 2004.
44. D. Ha, *et al., Tech. Dig. – Int. Electron Device Meet.* (2004). .
45. C. S. Park, *et al., IEEE Electron Device Lett.*, **25**, 372 (2004).
46. F. Ootsuka, *et al., Tech. Dig. – Int. Electron Device Meet.*, p. 575 (2000).
47. K. Ota, *et al., Tech. Dig. – Int. Electron Device Meet.*, p. 27 (2002).
48. T. Ghani, *et al., Tech. Dig. – Int. Electron Device Meet.*, p. 978 (2003).
49. C.-H. Ge, *et al., Tech. Dig. – Int. Electron Device Meet.*, p. 73 (2003).
50. M. Yang, *et al., Tech. Dig. – Int. Electron Device Meet.*, p. 453 (2003).
51. J. Kedzierski, *et al., Tech. Dig. – Int. Electron Device Meet.*, p. 57 (2000).
52. D. Connelly, *et al., IEEE Trans. Electron Devices*, **50**, 411 (2003).
53. M. Tao, *et al., Appl. Phys. Lett.*, **82**, 2593 (2003).
54. A. Yagishita, *et al.,* Extended Abstracts of the International Conference on Solid-State Devices and Materials, p. 708 (2003).
55. A. Kinoshita, *et al., Tech. Dig. – VLSI Sympos.*, p. 168 (2004).
56. N. Jin, *et al., IEEE Trans. Electron Devices*, **50**, 1876 (2003).
57. L. Esaki, *Phys. Rev.*, **109**, 603 (1958).
58. P. Mazumder, *et al., Proc. IEEE*, **86**, 664 (1998).
59. J. P. A. van der Wagt, *Proc. IEEE*, **87**, 571 (1999).

About the Author

Tsu-Jae King is a professor of Electrical Engineering and Computer Sciences at the University of California, Berkeley, where her research interests include nanoscale semiconductor devices and technology. She is presently on industrial leave of absence at Synopsys, Inc. (Mountain View, CA) leading the development of NDR device technology for compact, low-power memory applications. She can be reached at tking@eecs.berkeley.edu or tsujae@synopsys.com.