

INTERNATIONAL TECHNOLOGY ROADMAP FOR SEMICONDUCTORS

2003 EDITION

EMERGING RESEARCH DEVICES

THE ITRS IS DEvised AND INTENDED FOR TECHNOLOGY ASSESSMENT ONLY AND IS WITHOUT REGARD TO ANY COMMERCIAL CONSIDERATIONS PERTAINING TO INDIVIDUAL PRODUCTS OR EQUIPMENT.

TABLE OF CONTENTS

Emerging Research Devices	1
Scope	1
Difficult Challenges	1
Emerging Technology Sequence	2
Emerging Research Devices	3
Non-classical CMOS	3
Introduction.....	3
Non-classical CMOS—Definition and Discussion of Table Entries.....	3
Non-classical CMOS—An Emerging Device Technology Roadmap Scenario	9
Memory Devices	15
Introduction.....	15
Memory Taxonomy.....	15
Memory Devices—Definition and Discussion of Table Entries	19
Logic Devices	21
Introduction.....	21
Logic Devices—Definition and Discussion of Table Entries	27
Emerging Research Architectures	33
Introduction	33
Architectures—Definition and Discussion of Table Entries	33
Fine-Grained Parallel Implementations in Nanoscale Cellular Arrays	33
Defect Tolerant Architecture Implementations.....	34
Biologically Inspired Architecture Implementations	35
Coherent Quantum Computing.....	37
Emerging Technologies—A Functional Comparison	40
Introduction	40
Functional Parameterization and Comparison	40
Definition and Discussion of Table Entries.....	42
Emerging Technologies—A Critical Review	44
Introduction	44
Technologies beyond CMOS.....	44
Overall Technology Requirements	44
Charge-based Nanoscale Devices	45
Alternate Logic-State-Variable Nanoscale Devices	45
Potential Performance and Risk Assessment for Memory and Logic Devices	46
Relevance Criteria	46
Appendix MASTAR.....	49

LIST OF FIGURES

Figure 38	Emerging Technology Sequence	3
Figure 39	Estimation of Electrostatic Integrity (EI) for Bulk and Double-gate FETs	11
Figure 40	Impact of the Technology Boosters on HP, LOP, and LSTP CMOS Roadmaps in Terms of $I_{on}:I_{off}$ Ratio	12
Figure 41	Impact of the Technology Boosters on HP, LOP, and LSTP CMOS Roadmaps in Terms of Device Intrinsic Speed ($f=1/(CV/I)$).....	14
Figure 42	Parameterization of Emerging Technologies and CMOS— Speed, Size, Cost, and Switching Energy	41

LIST OF TABLES

Table 58	Emerging Technologies Difficult Challenges	2
Table 59a	Single-gate Non-classical CMOS Technologies	4
Table 59b	Multiple-gate Non-classical CMOS Technologies.....	5
Table 60	Technology Performance Boosters	10
Table 61	Memory Taxonomy	16
Table 62a	Emerging Research Memory Devices—Projected Parameters	17
Table 62b	Emerging Research Memory Devices—Experimental Parameters.....	18
Table 63a	Emerging Research Logic Devices—Projected Parameters.....	23
Table 63b	Emerging Research Logic Devices—Experimental Parameters	25
Table 64	Emerging Research Architecture Implementations.....	38
Table 65	Estimated Parameters for Emerging Research Devices and Technologies in the year 2016	42
Table 66	Technology Performance and Risk Evaluation for Emerging Research Memory Device Technologies (Potential/Risk)	48
Table 67	Technology Performance and Risk Evaluation for Emerging Research Logic Device Technologies (Potential/Risk)	48

EMERGING RESEARCH DEVICES

SCOPE

The quickening pace of MOSFET scaling is accelerating introduction of new technologies to extend CMOS beyond the 45 nm technology node. This acceleration simultaneously requires the industry to intensify research on two highly challenging thrusts. One is scaling CMOS into an increasingly difficult manufacturing domain well below the 90-nm node, and the other is an exciting opportunity to invent fundamentally new approaches to information and signal processing to sustain functional scaling beyond the domain of CMOS.

The primary goal of this section is to stimulate invention and research leading to feasibility demonstration for one or more Roadmap-extending concepts. This goal is accomplished by addressing the two technology-defining domains identified above—non-classical CMOS structures and memory technologies and completely new technological and architectural concepts for revolutionary Roadmap-extending information and signal processing applications. Technologies addressing CMOS scaling include both new materials and advanced MOSFET structures. The *Front End Processes* chapter discusses new materials required, for example, for the gate stack and for source/drain contacts. The *Process Integration, Devices, and Structures* chapter identifies technology requirements for CMOS structures to sustain performance and density scaling. Inclusion of a concept in this section does not in any way constitute advocacy or endorsement of that concept.

An important new theme of this section is to provide balanced technical assessments of leading approaches to non-classical CMOS device technologies and new information and signal processing approaches. Furthermore, the content has been expanded to provide additional quantitative depth necessary to compare projected and current performance of several emerging new technologies. The intent is two-fold. First is to “cast a broad net” to gather in one place substantive, alternative concepts for memory, logic, and information processing architectures that would, if successful, substantially extend the Roadmap beyond CMOS. As such, this discussion will provide a window into candidate approaches. Second is to provide a balanced, critical assessment of these emerging new device technologies for information processing. This broadened section, therefore, provides an industry perspective on emerging new device technologies and serves as a bridge between bulk CMOS and the realm of microelectronics beyond the end of CMOS scaling.

The discussion is divided into the following four categories: 1) Non-classical CMOS, 2) Memory Devices, 3) Logic Devices and 4) information processing Architectures. The discussions provide some detail regarding their operation principles, advantages, challenges, maturity, and current and projected performance. Also included is a preliminary but interesting comparison of the performance projections and cost attributes for several speculative new approaches to information and signal processing. An interesting observation of this comparison is that the emerging devices, technologies, and architectures, given their successful development, would extend applications of microelectronics to domains not accessible to CMOS, rather than competing directly with CMOS in the same domain.

DIFFICULT CHALLENGES

The microelectronics industry is facing two sets of difficult challenges related to extending integrated circuit technology to and beyond the end of CMOS scaling. One set of challenges relates to logic and the other relates to memory technologies. One difficult challenge related to logic in both the near- and the longer-term is to extend CMOS technology to and beyond the 45 nm node sustaining the historic annual increase of intrinsic speed of high-performance MPUs at 17%. This may require an unprecedented simultaneous introduction of two or more innovations to the device structure and/or gate-stack materials. Another longer-term challenge for logic is invention and reduction to practice of a new manufacturable information and signal processing technology addressing “beyond CMOS” applications. Solutions to the first may be critically important to extension of CMOS beyond the 45 nm node, and solutions to the latter could open opportunities for microelectronics beyond the end of CMOS scaling.

Another difficult challenge is the need of a new memory technology that combines the best features of current volatile and non-volatile memories in a fabrication technology compatible with CMOS process flow. This would provide a memory device fabrication technology required for both stand-alone and embedded memory applications. The ability of

2 Emerging Research Devices

an MPU to execute programs is limited by interaction between the processor and the memory, and scaling does not automatically solve this problem. The current evolutionary solution is to increase MPU cache memory, thereby increasing the floorspace that SRAM occupies on an MPU chip. This trend eventually leads to a decrease of the net information throughput. In addition, volatility of semiconductor memory requires external storage media with slow access (e.g., magnetic hard drives, optical CD, etc.). Therefore, development of *electrically accessible non-volatile* memory with *high speed* and *high density* would initiate a revolution in computer architecture. This development would provide a significant increase in information throughput even if traditional benefits of scaling were fully realized for nanoscale CMOS devices.

Table 58 Emerging Technologies Difficult Challenges

<i>Difficult Challenges ≥45 nm/Through 2009</i>	<i>Summary of Issues</i>
Implementation into manufacturing of non-classical MOSFET device structures integrated with new materials and processes (for example, a strained silicon channel integrated with a new high-κ gate dielectric material)	Selection of most promising device structure(s) and/or materials technologies ["Technology Booster(s)"] to sustain the required annual 17% increase in performance Introduction of two or more "Technology Booster" options (material, process, and/or device structure changes) simultaneously in a single node
Development and implementation into manufacturing of a non-volatile memory technology combining the best performance features of both volatile and non-volatile memory technologies for both stand-alone and embedded applications	Realization of a manufacturable, cost-effective fabrication technology for electrically accessible high-speed, high-density non-volatile RAM integrable with the fabrication process flow for CMOS logic
<i>Difficult Challenges <45 nm/Beyond 2009</i>	
Toward the end of CMOS scaling or beyond, discovery, reduction to practice, and implementation into manufacturing of novel, non-CMOS devices and architectures integrated (monolithically, mechanically, or functionally) with a CMOS platform technology	Discovery and reduction to practice of new information processing technologies integrable with silicon CMOS Discovery and reduction to practice of new, low-cost methods of manufacturing novel information processing technologies

EMERGING TECHNOLOGY SEQUENCE

Figure 38 shows an overview of the organization of the Emerging Research Devices section and illustrates the relationship of particular new concepts to the four functional categories that they each address—Non-classical CMOS, Memory, Logic, and Architectures. A category for Architectures is included to emphasize the point that because both new systems architectures and new device technologies will drive development of the other, synergistic/collaborative development of the two together can be very rewarding. This figure illustrates one simplified example of a richly diverse set of emerging application-specific concepts and technologies addressing different functions. It includes several highly speculative approaches. Many of these concepts likely will not mature to manufacturing or application. The important message here is that the emergence of many new ideas and technologies, several of which are suitable for only certain function(s) and do not have broad application, may be signaling a coming dispersion of microelectronics technologies to address an increasingly diverse set of market-driven applications. Integration of Systems on Chip (SoC) and in a package (SiP) at low cost and within a prescribed form factor will undoubtedly continue use of CMOS as the functional integration platform. This confluence of functions drives the need to integrate dissimilar technologies and functions in a high-performance, low-cost fashion with CMOS platforms.

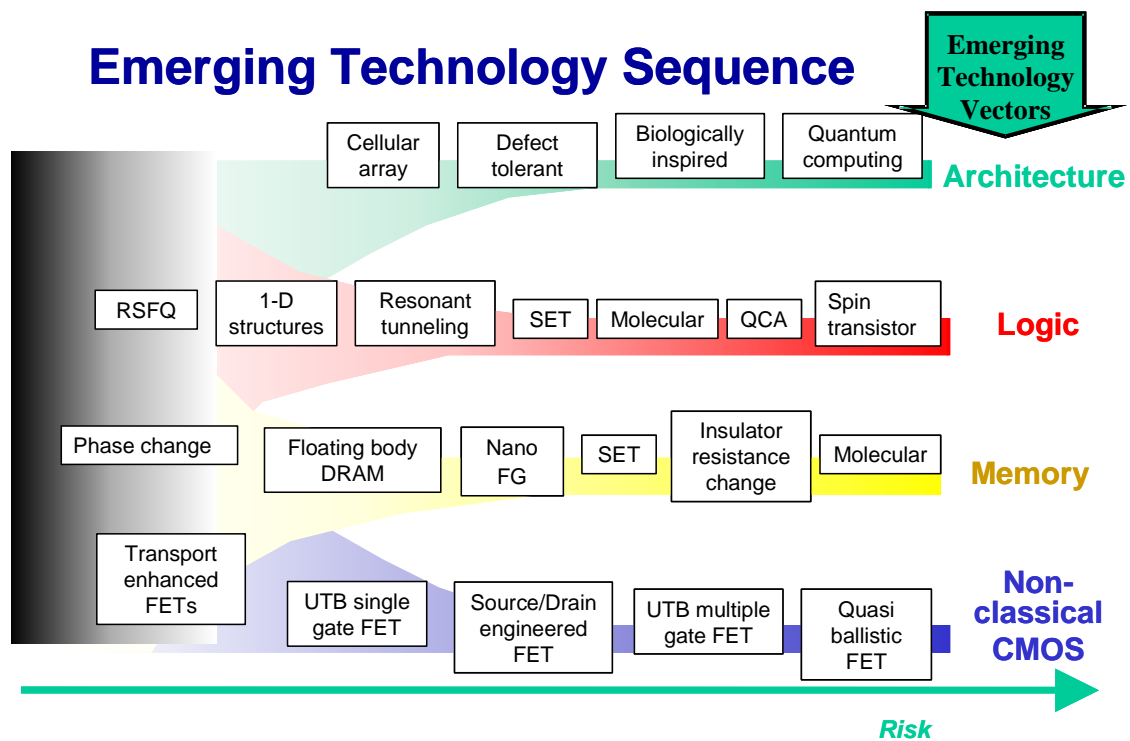


Figure 38 Emerging Technology Sequence

EMERGING RESEARCH DEVICES

NON-CLASSICAL CMOS

INTRODUCTION

Non-classical CMOS includes those advanced MOSFETs, shown in Tables 59a and 59b, which provide a path to scaling CMOS to the end of the Roadmap using new transistor structural designs and new materials. For digital applications, the scaling challenges include controlling leakage currents and short-channel effects; increasing saturation current while reducing the power supply; control of device parameters (e.g., threshold voltage, leakage) across the chip and from chip to chip. For analog/mixed-signal/RF applications, the challenges additionally include sustaining linearity, low noise figure, power-added-efficiency, and transistor matching. The industry and academic communities are pursuing two avenues to meeting these challenges—new transistor structures and new materials. New transistor structures seek to improve the electrostatics of the MOSFET; provide a platform for introduction of new materials; and accommodate the integration needs of new materials. New materials include those used in the gate stack (high- κ dielectric and electrode materials), those used in the conducting channel that have improved carrier transport properties, as well as new materials used in the source/drain regions with reduced resistance and carrier injection properties. Additionally, the combination of new device structures and new materials enables new operating principles that may provide new behavior and functionality beyond the constraints of bulk planar or classical CMOS.

NON-CLASSICAL CMOS—DEFINITION AND DISCUSSION OF TABLE ENTRIES

Transport-enhanced FETs—Improvements in transistor drive current for improved circuit performance can be achieved by enhancing the average velocity of carriers in the channel. Approaches to enhancing transport include mechanically straining the channel layer to enhance carrier mobility and saturation velocity, and employing alternative channel materials such as silicon-germanium, germanium, or III-V compound semiconductors with electron and hole mobilities and carrier velocities higher than those in silicon. A judicious choice of crystal orientation and current transport direction may also provide transport enhancement¹. However, an important issue is how to fabricate transport enhanced channel layers (such as a strained Si layer) in several of the non-classical CMOS transistor structures (e.g., the multiple gate structures discussed in Table 59b).

¹ S. Takagi, "Re-examination of Sub-band Structure Engineering in Ultra-short Channel MOSFETs under Ballistic Carrier Transport," VLSI Technology Symposium (2003) 115.

4 Emerging Research Devices

Table 59a Single-gate Non-classical CMOS Technologies

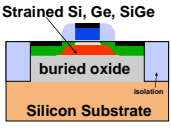
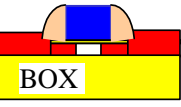
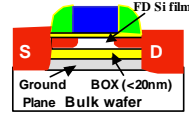
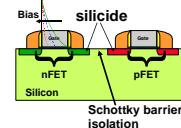
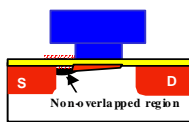
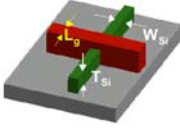
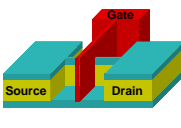
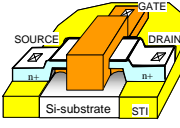
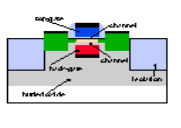
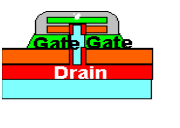
Device	Transport-enhanced FETs	Ultra-thin Body SOI FETs		Source/Drain Engineered FETs	
					
Concept	Strained Si, Ge, SiGe, SiGeC or other semiconductor; on bulk or SOI	Fully depleted SOI with body thinner than 10 nm	Ultra-thin channel and localized ultra-thin BOX	Schottky source/drain	Non-overlapped S/D extensions on bulk, SOI, or DG devices
Application/Driver	HP CMOS	HP, LOP, and LSTP CMOS	HP, LOP, and LSTP CMOS	HP CMOS	HP, LOP, and LSTP CMOS
Advantages	<ul style="list-style-type: none"> High mobility 	<ul style="list-style-type: none"> Improved subthreshold slope No floating body Potentially lower E_{eff} 	<ul style="list-style-type: none"> SOI-like structure on bulk Shallow junction by geometry Junction silicidation as on bulk Improved S-slope and SCE 	<ul style="list-style-type: none"> Low source/drain resistance 	<ul style="list-style-type: none"> Reduced SCE and DIBL Reduced parasitic gate capacitance
Particular Strength	<ul style="list-style-type: none"> High mobility without change in device architecture 	<ul style="list-style-type: none"> Low diode leakage Low junction capacitance No significant change in design with respect to bulk 	<ul style="list-style-type: none"> Quasi-DG operation due to ground plane effect enabled by the ultra thin BOX Bulk compatible 	<ul style="list-style-type: none"> No need for abrupt S/D doping or activation 	<ul style="list-style-type: none"> Very low gate capacitance
Potential Weakness	<ul style="list-style-type: none"> Material defects and diode leakage (only for bulk) Process compatibility and thermal budget Operating temperature 	<ul style="list-style-type: none"> Very thin silicon required with low defect density V_{th} adjustment difficult Selective epi required for elevated S/D 	<ul style="list-style-type: none"> Ground plane capacitance Selective epi required for channel and S/D 	<ul style="list-style-type: none"> Ultra-thin SOI required NFET silicide material not readily available Parasitic potential barrier 	<ul style="list-style-type: none"> High source/drain resistance Reliability Advantageous only for very short devices
Scaling Issues	Bandgap usually smaller than Si	Control of Si film thickness	Process becomes easier with L_g down-scaling (shorter tunnel)	No particular scaling issue	Sensitivity to L_g variation
Design Challenges	Compact model needed	None	None	Compact model needed	Compact model needed
Gain/Loss in Layout compared to Bulk	No difference	No difference	No difference	No difference	No difference
Impact on I_{on}/I_{off} compared to Bulk	<ul style="list-style-type: none"> Improved by 20–30% (from MASTAR supposing $\mu_{eff} \times 2$) 	<ul style="list-style-type: none"> Improved by 15–20% (from MASTAR supposing $E_{eff}/2$ and $S=75mV/dec$) 	<ul style="list-style-type: none"> Improved by 15–20% (from MASTAR supposing $E_{eff}/2$ and $S=75mV/dec$) 	<ul style="list-style-type: none"> Improved by 10–15% (from MASTAR supposing $R_{series}=0$) 	<ul style="list-style-type: none"> Both shifted to lower values
Impact on CV/I compared to Bulk	<ul style="list-style-type: none"> Lowered by 15–20% (from MASTAR supposing $\mu_{eff} \times 2$) 	<ul style="list-style-type: none"> Lowered by 10–15% (from MASTAR supposing $E_{eff}/2$ and $S=75mV/dec$) 	<ul style="list-style-type: none"> Lowered by 10–15% (from MASTAR supposing $E_{eff}/2$ and $S=75mV/dec$) 	<ul style="list-style-type: none"> Lowered by 10–15% (from MASTAR supposing $R_{series}=0$) 	<ul style="list-style-type: none"> Constancy or gain due to lower gate capacitance
Analog Suitability G_m/G_d advantage compared to Bulk	Not clear	Potential for slight improvement	Potential for slight improvement	Not clear	Not clear

Table 59b Multiple-gate Non-classical CMOS Technologies

Device	Multiple Gate FETs				
	N-Gate ($N > 2$) FETs	Double-gate FETs			
					
Concept	Tied gates (number of channels > 2)	Tied gates, side-wall conduction	Tied gates planar conduction	Independently switched gates, planar conduction	Vertical conduction
Application/Driver	HP, LOP, and LSTP CMOS	HP, LOP, and LSTP CMOS	HP, LOP, and LSTP CMOS	LOP and LSTP CMOS	HP, LOP, and LSTP CMOS
Advantages	<ul style="list-style-type: none"> Higher drive current $2 \times$ thicker fin allowed 	<ul style="list-style-type: none"> Higher drive current Improved subthreshold slope Improved short channel effect 	<ul style="list-style-type: none"> Higher drive current Improved subthreshold slope Improved short channel effect 	<ul style="list-style-type: none"> Improved short channel effect 	<ul style="list-style-type: none"> Potential for 3D integration
Particular Strength	<ul style="list-style-type: none"> Thicker Si body possible 	<ul style="list-style-type: none"> Relatively easy process integration 	<ul style="list-style-type: none"> Process compatible with bulk and on bulk wafers Very good control of silicon film thickness 	<ul style="list-style-type: none"> Electrically (statically or dynamically) adjustable threshold voltage 	<ul style="list-style-type: none"> Lithography independent L_g
Potential weakness	<ul style="list-style-type: none"> Limited device width Corner effect 	<ul style="list-style-type: none"> Fin thickness less than the gate length Fin shape and aspect ratio 	<ul style="list-style-type: none"> Width limited to $< 1 \mu\text{m}$ 	<ul style="list-style-type: none"> Difficult integration Back-gate capacitance Degraded subthreshold slope 	<ul style="list-style-type: none"> Junction profiling difficult Process integration difficult Parasitic capacitance Single gate length
Scaling Issues	<ul style="list-style-type: none"> Sub-lithographic fin thickness required 	<ul style="list-style-type: none"> Sub-lithographic fin thickness required 	<ul style="list-style-type: none"> Bottom gate larger than top gate 	<ul style="list-style-type: none"> Gate alignment 	<ul style="list-style-type: none"> Si vertical channel film thickness
Design Challenges	<ul style="list-style-type: none"> Fin width discretization 	<ul style="list-style-type: none"> Fin width discretization 	<ul style="list-style-type: none"> Modified layout 	<ul style="list-style-type: none"> New device layout 	<ul style="list-style-type: none"> New device layout
Gain/Loss in Layout compared to Bulk	<ul style="list-style-type: none"> No difference 	<ul style="list-style-type: none"> No difference 	<ul style="list-style-type: none"> No difference 	<ul style="list-style-type: none"> No difference 	<ul style="list-style-type: none"> Up to 30% gain in layout density
Advantage in I_{on}/I_{off} compared to Bulk	<ul style="list-style-type: none"> Improved by 20–30% (from MASTAR assuming $E_{eff}/2$ and $S=65\text{V/decade}$) 	<ul style="list-style-type: none"> Improved by 20–30% (from MASTAR assuming $E_{eff}/2$ and $S=65\text{V/decade}$) 	<ul style="list-style-type: none"> Improved by 20–30% (from MASTAR assuming $E_{eff}/2$ and $S=65\text{V/decade}$) 	<ul style="list-style-type: none"> Potential for improvement 	<ul style="list-style-type: none"> Improved by 20–30% (from MASTAR assuming $E_{eff}/2$ and $S=65\text{V/decade}$)
Advantage in CV/I compared to Bulk	<ul style="list-style-type: none"> Lowered by 15–20% (from MASTAR assuming $E_{eff}/2$ and $S=65\text{V/decade}$) 	<ul style="list-style-type: none"> Lowered by 15–20% (from MASTAR assuming $E_{eff}/2$ and $S=65\text{V/decade}$) 	<ul style="list-style-type: none"> Lowered by 15–20% (from MASTAR assuming $E_{eff}/2$ and $S=65\text{V/decade}$) 	<ul style="list-style-type: none"> Potential for improvement 	<ul style="list-style-type: none"> Lowered by 15–20% (from MASTAR assuming $E_{eff}/2$ and $S=65\text{V/decade}$)
Analog Suitability G_m/G_d advantage compared to Bulk	<ul style="list-style-type: none"> Potential for improvement 	<ul style="list-style-type: none"> Potential for improvement 	<ul style="list-style-type: none"> Potential for improvement 	<ul style="list-style-type: none"> Potential for improvement 	<ul style="list-style-type: none"> Potential for improvement

6 Emerging Research Devices

References for Table 59a:

Transport-enhanced FETs / Strained Si, Ge, SiGe, SiGeC or other semiconductor; on bulk or SOI / HP CMOS:

- C. Chiu, "A sub-400 Degree C Germanium MOSFET Technology with High-k Dielectric and Metal Gate," *IEDM* (2002), 437–440.
- H. Shang, "High Mobility p-Channel Germanium MOSFETs with a Thin Ge Oxyntiride Gate Dielectric," *IEDM* (2002), San Francisco, California.
- C.W. Leitz, "Hole Mobility Enhancements in Strained Si/Si/sub 1-y/Ge/sub y/ p-Type Metal-oxide-semiconductor Field-effect Transistors Grown on Relaxed Si/sub 1-x/Ge/sub x/ (x<y) Virtual Substrates," *Applied Physics Letters*, Vol. 79, No. 25, December 17, 2001.
- M. Lee, "Strained Ge Channel p-Type Metal-oxide-semiconductor Field-effect Transistor Grown on Si x Ge 1-x / Si Virtual Substrates," *Applied Physics Letters*, Vol. 79, No. 20, November 21, 2001.
- B.H. Lee, "Performance Enhancement on Sub-70 nm Strained Silicon SOI MOSFETs on Ultra Thin Thermally Mixed Strained Silicon/SiGe on Insulator (TM-SGOI) Substrate with Raised S/D," *IEDM* (December 11, 2002), San Francisco, California .
- T. Mizuno, "High Performance CMOS Operation of Strained-SOI MOSFETs Using Thin Film SiGe-on-Insulator Substrate," *VLSI Technology Symposium* (June 11–13, 2002), Honolulu, Hawaii.
- T. Tezuka, "High-performance Strained Si-on-Insulator MOSFETs by Novel Fabrication Processes Utilizing Ge-Condensation Technique," *VLSI Technology Symposium* (June 11–13, 2002), Honolulu, Hawaii.
- N. Collaert, "High-Performance Strained Si/SiGe pMOS Devices With Multiple Quantum Wells," *IEEE Trans. Nanotechnology*, Vol. 1, No. 4, December 2002, 190–194.
- T. Ernst, "A New Si:C Epitaxial Channel nMOSFET Architecture with Improved Drivability and Short-channel Characteristics," *VLSI Technology Symposium*, (June 10-12, 2003), Kyoto, Japan.
- Qi Xiang, "Strained Silicon NMOS with Nickel-Silicide Metal Gate," *VLSI Technology Symposium* (June 10-12, 2003), Kyoto, Japan.
- J.R. Hwang, "Performance of 70 nm Strained-Silicon CMOS Devices," *VLSI Technology Symposium* (June 10-12, 2003), Kyoto, Japan.
- T. Mizuno, "(110)-Surface Strained-SOI CMOS Devices with Higher Carrier Mobility," *VLSI Technology Symposium* (June 10-12, 2003), Kyoto, Japan.
- C.H. Huang, "Very Low Defects and High-performance Ge-On-Insulator p-MOSFETs with Al₂O₃ Gate Dielectrics," *VLSI Technology Symposium* (June 10-12, 2003), Kyoto, Japan.

Ultra-thin Body SOI FETs / Fully depleted SOI with body thinner than 10 nm / HP, LOP, and LSTP CMOS:

- B. Doris, "Extreme Scaling with Ultra-thin Si Channel MOSFETs," *IEDM* (December 8–11, 2002), San Francisco, California, 267–270.
- R. Chau, "A 50 nm Depleted-Substrate CMOS Transistor (DST)," *IEDM* (December 2–5, 2001), Washington, D.C, 621–624.
- H. VanMeer, "70 nm Fully-Depleted SOI CMOS Using a New Fabrication Scheme: The Spacer/Replacer Scheme," *VLSI Symposium* (June 11–13, 2002), Honolulu, Hawaii.
- T. Schultz, "Impact of Technology Parameters on Inverter Delay of UTB-SOI CMOS," *SOI Conference*, (October 7–10, 2002), Williamsburg, Virginia, 176–178.
- A. Vandoreen, "Ultra-thin Body Fully-depleted SOI Devices with Metal Gate (TaSiN) Gate, High k (HfO₂) Dielectric and Elevated Source/Drain Extensions," *SOI Conference*, (October 7–10, 2002), Williamsburg, Virginia, 205–206.
- B. Yu, "Scaling Towards 35 nm Gate Length CMOS," *VLSI Symposium* (June 12–14, 2001), Kyoto, Japan, 9–10.
- Y.K. Choi, "Ultra-thin Body PMOSFETs with Selectively Deposited Ge Source/Drain," *VLSI symposium* (June 12–14, 2001), Kyoto, Japan, 19–20.
- K. Uchida, "Experimental Study on Carrier Transport Mechanism in Ultrathin-body SOI n and p MOSFETs with SOI Thickness Less Than 5 nm," *IEDM* (December 8–11, 2002), San Francisco, California, 47–50.

Ultra-thin Body SOI FETs / Ultra-thin channel and localized ultra-thin BOX / HP, LOP, and LSTP CMOS:

- M. Jurczak, "SON (Silicon On Nothing) – A New Device Architecture for the ULSI Era," *Symp. VLSI Technology Proceedings*, (June 1999), 29–30.
- T. Skotnicki, "Heavily Doped and Extremely Shallow Junctions on Insulator – by SONCTION (SilicON Cut-off Junction) Process," *IEDM*, (December 1999), 513–516.
- M. Jurczak, "SON (Silicon On Nothing) – An Innovative Process for Advanced CMOS," *IEEE Transactions on Electron Devices*, (November 2000), 2179.
- S. Monfray, "First 80 nm SON (Silicon-On-Nothing) MOSFETs with Perfect Morphology and High Electrical Performance," *IEDM*, (December 2001), 645–648.
- S. Monfray, "SON (Silicon-On-Nothing) P-MOSFETs with Totally Silicided (CoSi₂) Polysilicon on 5 nm-Thick Si-films: The Simplest Way to Integration of Metal Gates on Thin FD Channels," *IEDM*, (December 2002), 263.
- S. Monfray, "Highly-performant 38 nm SON (Silicon-On-Nothing) P-MOSFETs with 9 nm-Thick Channels," *IEEE SOI Conference Proceedings*, (October 2002), 20.
- T. Sato, "SON (Silicon On Nothing) MOSFET using ESS (Empty Space in Silicon) Technique for SoC Applications," *IEDM*, (December 2001), 809.

Source/Drain Engineered FETs / Schottky source/drain / HP CMOS

- J. Kedzierski, "Complementary Silicide Source/Drain Thin-body MOSFETs for the 20 nm Gate Length Regime," *IEDM (December 2002)* San Francisco, California.
- R. Rishon, "New Complementary Metal-oxide Semiconductor Technology with Self-aligned Schottky Source/Drain and Low-resistance T-gates," *Journal of Vacuum Science Technology*, 1997, 2795–2798.
- J.P. Snyder, "Experimental Investigation of a PtSi Source and Drain Field Emission Transistor," *Applied Physics Letters*, Vol. 67, No. 10, September 4, 1995.

Source/Drain Engineered FETs // Non-overlapped S/D Extensions on Bulk, SOI, or DG devices // HP, LOP, and LSTP CMOS.

- F. Boeuf, "16 nm Planar NMOSFET Manufacturable within State-of-the-art CMOS Process Thanks to Specific Design and Optimization," *IEDM (December 2001)*, Washington, D.C., 637–640.
- H. Lee, "DC and AC Characteristics of Sub-50-nm MOSFETs with Source/Drain-to-gate Nonoverlapped Structure," *IEEE Trans. Nanotechnology*, Vol. 1, No. 4, December 2002, 219–225.

References for Table 59b:**Multiple-gate FETs / N-Gate (N>2) FET / Tied gates (number of channels >2) / HP, LOP, and LSTP CMOS**

- R. Chau, "Advanced Depleted Substrate Transistor: Single-gate, Double-gate, and Tri-gate," *Solid State Device Meeting (2002)*, 68-69.
- Fu-Liang Yang, "25 nm CMOS Omega FETs," *IEDM (December 2002)*, 255.
- J. Colinge, "Silicon-on-insulator Gate-all-around Device," *IEDM (December 1990)*, 595.
- B. Doyle, "Tri-gate Fully-depleted CMOS Transistors Fabrication, Design and Layout," *VLSI (June 2003)*, 133.
- Z. Krivokapic, "High Performance 45 nm CMOS Technology with 20 nm Multi-gate Devices," *SSDM (September 2003)*, 760.

Multiple-gate FETs / Double-gate FET / Tied-gates, side-wall conduction / HP, LOP, and LSTP CMOS

- Y.K. Choi, "FinFET Process Refinements for Improved Mobility and Gate Work Function Engineering," *IEDM (December 2002)*, 259.
- J. Kedzierski, "Metal-gate FinFET and Fully-depleted SOI Devices Using Total Gate Silicidation," *IEDM (December 2002)*, 247.
- B. Yu, "FinFET Scaling to 10 nm Gate Length," *IEDM (December 2002)*, 251.
- T. Park, "Fabrication of Body-Tied FinFETs (Omega MOSFETs) Using Bulk Si Wafers," *VLSI (June 2003)*, 135.

Multiple-gate FETs / Double-gate FET / Tied-gates, planar conduction / HP, LOP, and LSTP CMOS

- S. Monfray, "50 nm – Gate All Around (GAA) – Silicon On Nothing (SON) – Devices: A Simple Way to Co-integration of GAA Transistors with Bulk MOSFET Process," *VLSI (June 2002)*, 108.
- Lee, "A Manufacturable Multiple Gate Oxynitride Thickness Technology for System on a Chip," *IEDM (December 1999)*, 71.
- H.S.P. Wong, "Self Aligned (top and bottom) Double-Gate MOSFET with a 25 nm Thick Silicon Channel," *IEDM (December 1997)*, 427.
- G. Neudeck, "Novel Silicon Epitaxy for Advanced MOSFET Devices," *IEDM (December 2000)*, 169.
- S.M. Kim, "A Novel MBC (Multi-bridge-channel) MOSFET: Fabrication Technologies and Characteristics," *Si-Nanoworkshop (2003)*, 18.

Multiple-gate FETs / Double-gate FET / Independently switched gates, planar conduction / LOP and LSTP CMOS.

- I. Yang, "IEEE Transactions of Electron Devices," (1997), 822.
- K.W. Guarini, "Triple-self-aligned, Planar Double-Gate MOSFETs: Devices and Circuits," *IEDM (December 2001)*, 425.

Multiple-gate FETs / Double-gate FET / Vertical conduction / HP, LOP, and LSTP CMOS

- J.M.Hergenrother, "The Vertical Replacement-gate (VRG) MOSFET: a 50-nm vertical MOSFET with Lithography-independent Gate Length," *IEDM (December 1999)*, 75.
- J.M. Hergenrother, "50 nm Vertical Replacement-gate (VRG) nMOSFETs with ALD HfO₂ and Al₂O₃ Gate Dielectrics," *IEDM (December 2001)*, 51–54.
- E. Josse, "High Performance 40 nm Vertical MOSFET within a Conventional CMOS Process Flow," *VLSI (June 2001)*, 55–56.
- P. Verheyen, "A 50 nm Vertical Si/sub 0.70/Ge/sub 0.30/Si/sub 0.85/Ge/sub 0.15/ pMOSFET with an Oxide/nitride Gate Dielectric," *Conference: 2001 International Symposium on VLSI Technology, Systems, and Applications. Proceedings of Technical Papers (Cat. No.01TH8517)*, IMEC, Leuven, Belgium, 15–18
- B. Goebel, "Fully Depleted Surrounding Gate Transistor (SGT) for 70 nm DRAM and Beyond," *IEDM (December 2002)*, 275.
- Meishoku Masahara, "15-nm-Thick Si Channel Wall Vertical Double-Gate MOSFET," *IEDM (December 2002)*, 949.

8 Emerging Research Devices

Ultra-thin-body SOI FETs—A very thin transistor body is employed to ensure good electrostatic control of the channel by the gate in the “off” state. Typically, the ratio of the channel length to the channel thickness will be ≥ 3 . Hence an extremely thin (< 4 nm) Si channel is required to scale CMOS to the 22 nm node. The use of a lightly doped or undoped body provides immunity to V_t variations due to statistical dopant fluctuations, as well as enhanced carrier mobilities for higher transistor drive current. The *localized and ultra-thin BOX FET* is an UTB SOI-like FET in which a thin Si channel is locally isolated from the bulk-Si substrate by a thin (10–30 nm) buried dielectric layer. This structure combines the best features of the classical MOSFET (e.g., deep source/drain contact regions for low parasitic resistance) with the best features of SOI technology (improved electrostatics). The increased capacitive coupling between the source, drain, and channel with the conducting substrate through the ultra-thin BOX has the potential of reducing the speed of the device but also of improving the electrostatic integrity of the device. The former may be traded against the latter (by reducing the channel doping) that eventually leads to moderately improved speed for a constant I_{off} .

Source/drain engineered FETs—Engineering the source/drain is becoming critically important to maintaining the source and drain resistance to be a reasonable fraction ($\sim 10\%$) of the channel resistance. Two sub-category structures are described for providing engineered source/drain structures. First is the *Schottky source/drain* structure. In this case, the use of metallic source and drain electrodes minimizes parasitic series resistance and eliminates the need for ultra-shallow p–n junctions. Metals or silicides which form low (near zero) Schottky barrier heights in contact with silicon (i.e., a low-work-function metal for NMOS, and a high-work-function metal for PMOS) are required to minimize contact resistance and maximize transistor drive current in the “on” state. An ultra-thin body is needed to provide low leakage in the “off” state. Second is the *reduced fringing/overlap gate FET*. As MOSFET scaling continues, the parasitic capacitance between the gate and source/drain detrimentally affects circuit performance and its impact becomes more significant as the gate length is scaled down. For gate lengths below ~ 20 nm, transistor optimization for peak circuit performance within leakage current constraints will likely dictate a structure wherein the gate electrode does not overlap the source or drain to minimize the effect of parasitic fringing/overlap capacitance. Due to lengthening of its electrical channel, the non-overlapped gate structure does not require ultra-shallow source/drain junctions in order to provide good control of short-channel effects. Also, the increase of source/drain resistance usually expected for the non-overlap transistor is reduced with decreasing gate length, thus providing a new optimization paradigm for extremely short devices.

N-gate ($N > 2$) FETs—In the N-gate MOSFET current flows horizontally (parallel to the plane of the substrate) between the source and drain along vertical channel surfaces, as well as one or more horizontal channel surfaces. The large number of gates provides for improved electrostatic control of the channel, so that the Si body thickness and width can be larger than for the ultra-thin-body SOI and double-gate FET structures, respectively. The gate electrodes are formed from a single deposited gate layer and are defined lithographically. They are tied together electrically and are self-aligned with each other as well as the source/drain regions. The principal advantage of the structure resides in the relaxation of the needs on the thinness of the Si-body or the vertical fin. The challenge is in slightly poorer electrostatic integrity than with double-gate structures.

Double-gate FETs—A variety of double-gate MOSFET structures have been proposed to further improve engineering of the channel electrostatics and, in some cases, to provide independent control of two isolated gates for low-power and, perhaps, mixed-signal applications. Four typical double-gate structures are described in this section. First is the *tied double-gate, sidewall conduction structure*. This is a double-gate transistor structure in which current flows horizontally (parallel to the plane of the substrate) between the source and drain, along opposite vertical channel surfaces. The width of the vertical silicon fin is narrow (smaller than the channel length) to provide adequate control of short-channel effects. A lithographically defined gate straddles the fin, forming self-aligned, electrically connected gate electrodes along the sidewalls of the fin. The principal advantage with this structure is the planar bulk-like layout and process. The major challenge is with fabrication of thin fins that need to be a fraction ($\frac{1}{3}$ – $\frac{1}{2}$) of the gate length thus requiring sub-lithographic techniques.

The second structure is the *tied double-gate planar FET*. In this structure, current flows horizontally (parallel to the plane of the substrate) between the source and drain along opposite horizontal channel surfaces. The top and bottom gate electrodes are deposited in the same step and are defined lithographically. They may or may not be self-aligned to each other, and are electrically connected to one another. The source/drain regions are typically self-aligned to the top gate electrode. The principal advantages of this structure reside in the simplicity of the process (closest to bulk planar process) and in the compactness of the layout (same as for bulk planar) as well as in its compatibility with bulk layout (no need for redesigning libraries). Also important is that the channel thickness is determined by epitaxy, rather than etching, and thus is very well controlled. The challenge resides in the doping of the poly in the bottom gate (shadowed by the channel), but this problem disappears automatically when switching to a metal-like gate electrode. Another challenge is in the fabrication process, particularly for those structures requiring alignment of the top and bottom gate electrodes.

The third structure is the *independently switched double-gate (ground-plane) FET*. This structure is similar to the planar tied double-gate FET, except that the top and bottom gate electrodes are electrically isolated to provide for independent biasing of the two gates. The top gate is typically used to switch the transistor “on” and “off,” while the bottom gate is used for dynamic (or static) V_t adjustment. The principal advantage is in the very low I_{off} this structure offers. The disadvantage is in rather poor subthreshold behavior and in the relaxed layout.

The fourth structure is the *vertical transistor*. In this case, current flows between the source and drain in the vertical direction (orthogonal to the plane of the substrate) along two or more vertical channel surfaces. The gate length, hence the channel length, is defined by the thickness of the single deposited gate layer, rather than by a lithographic step. The gate electrodes are electrically connected, and are vertically self-aligned with each other and with the diffused source/drain extension regions. The principal advantage with this structure is that the channel length is defined by epitaxy rather than by lithography (possibility of very short and well-controlled channels). The disadvantage is this structure requires a challenging process and the layout is different from that for bulk transistors.

NON-CLASSICAL CMOS—AN EMERGING DEVICE TECHNOLOGY ROADMAP SCENARIO

Introduction—As investments relative to the majority of the non-classical CMOS structures presented above may be very large, it would be quite helpful to assess the gain in performance they promise. This knowledge will likely contribute to the technical justification and validity of the strategic R&D decisions that will be required to develop and implement one or more of these options. For many reasons this is a very difficult task. First, the properties of new materials may provide some surprises. As one example, our knowledge of these material properties is often based on isolated large volume samples, whereas in CMOS applications very thin and low volume layers are most common. Second, integration of these materials into a CMOS process may reveal undesirable interactions and place these materials under mechanical stress or lead to their inter-diffusion, etc, that may alter their properties. Third, the physics of new device structures is not always completely understood. Lastly, even the validity of numerical simulation results and tools are subject to debate, sometimes leading to large discrepancies depending on the choice of tools, models, and parameters. Frequently, a new structure or material gives mediocre results from first attempts at integration, thus precluding the possibility of calibration of simulation tools and of experimental verification of predictions. Years of difficult R&D efforts are sometimes necessary to prove the real value of a technological innovation.

Given the strategic importance of this task, an example of one possible emerging device architecture roadmap scenario is offered and discussed. Considering the precautions and uncertainties discussed above, qualitative guidelines and relative estimations are sought rather than quantitative accuracy.

The methodology employed for this task consists in using simple and widely recognized analytical expressions describing the conventional planar MOSFET physics. A set of equations (called MASTAR^{2,3}) served as a backup to the Excel spreadsheet used for the development of the logic technology requirements tables in the PIDS section [[link to PIDS chapter](#)]. The main equations have been aligned and calibrated between both tools, so as to ensure very close agreement for all three PIDS technology tables (HP, LOP, and LSTP). The methodology used in the spreadsheet model to assemble the PIDS technology requirements tables consists in satisfying the intrinsic speed $(CV/I)^{-1}$ improvement rate (17% per year) by requiring the necessary values of I_{on} (transistor “on”-current) but without linking these requirements to a given technological realization⁴. In contrast, the following analysis is aimed at finding this link and at assessing the magnitude of improvement of the entries presented in the non-classical CMOS Tables 59a and 59b.

In order to do so, a table of modifications was established entitled “Technology Performance Boosters,” given in Table 60. These modifications used in the MASTAR equations allow rough estimations of the performance gains in terms of I_{on} , C_{gate} , and I_{off} . ([A complete description of this method is provided in the supplemental section.](#)) Therefore, in addition to the precautions due to new materials and structures, one needs to be aware that the employed methodology cannot give more than a first order estimate. The effect of the Technology Performance Boosters is discussed on electrostatic integrity of the device, on the $I_{\text{on}}-I_{\text{off}}$ ratio, and on the CV/I .

² The MASTAR executable code file along with the User’s Guide are available as part of the ITRS 2003 background documentation. Refer to the Appendix of this section for instructions on downloading.

³ T. Skotnicki and F. Boeuf, “CMOS Technology Roadmap—Approaching Up-Hill Specials,” in *Ninth International Symposium on Silicon Materials Science and Technology, Process Integration, ECS 2002*.

⁴ Nonetheless, the required I resulting from the $(CV/I)^{-1}$ is matched with the I_{on} value resulting from the spread-sheet model (very close to MASTAR) in which some parameters are boosted to account for new materials and novel device structures in an implicit way (without making any direct link between those two). Such an approach is believed to help the reliability of predictions. The values of the boosters were agreed between the PIDS and ERD working groups, but their nature was left to be established through the more in-depth analysis carried out by the ERD group (this non-classical CMOS architectures section summarizes the results of this analysis).

Table 60 Technology Performance Boosters

Technology Performance Boosters				
Nature	Translation for I_{on}	Translation for C_{gate}	Translation for I_{off}	MASTAR Default Value
Strained-Si, Ge, etc.	$\mu_{eff} \times B_{mob}$	NA	NA	Strained-Si $B_{mob}=2$
Ultra-thin Body (Single Gate)	$E_{eff} \times B_{field}$ and $d \times B_d$	NA	S=75mV/decade and $X_j=T_{dep}=T_{si}$	$B_{field}=0.5$ $B_d=0.5$
Metal Gate/ High- κ Gate Dielectric	$T_{ox_el} - B_{gate}$	$T_{ox_el} - B_{gate}$	$T_{ox_el} - B_{gate}$	$B_{gate} =$ 4A NMOS
Ultra-thin Body (Double Gate)	$E_{eff} \times B_{field}$ and $d \times B_d$	NA	S=65mV/decade and $X_j=T_{dep}=T_{si}/2$	$B_{field}=0.5$ $B_d=0$
Ballistic	$V_{sat} \times (B_{ball})$	NA	NA	$B_{ball}=1.3$
Reduced Gate Parasitic Capacitance (Fringing and/or Overlap)	NA	$C_{fringe} \times B_{fring}$	NA	$B_{fring}=0.5$
Metallic S/D Junction	$R_{sd} \times B_{junc}$	NA	NA	$B_{junc}=0.5$

The boosters used in Table 60 are defined as follows:

B_{mob} —the effective mobility (μ_{eff}) improvement factor (long channel mobility) used for example to account for strained-Si channel material

B_{field} —the effective field (E_{eff}) reduction factor used to account for lower effective field (and thus higher mobility) in UTB devices

B_{gate} —the reduction in the effective electrical oxide thickness in inversion (T_{ox_el}) accounting for cancellation of the polydepletion effect and thus used to account for a metallic gate.

B_d —the body effect coefficient (d) reduction factor used to account for smaller d in UTB devices

B_{ball} —the saturation velocity (v_{sat}) effective improvement factor used to account (artificially) for a (quasi-) ballistic transport

B_{fring} —the fringing capacitance (C_{fring}) reduction factor used to account for reduced fringing capacitance

B_{junc} —the series resistance (R_{sd}) reduction factor used for example to account for metallic (Schottky) junctions

Sustaining the electrostatic integrity of ultra-scaled CMOS—The electrostatic integrity (EI) of a device reflects its resistance to parasitic 2D effects such as SCE and DIBL. SCE is defined as the difference in threshold voltage between long-channel and short-channel FETs measured using small V_{ds} . DIBL is defined as the difference in V_t measured for short-channel FETs using a small and a nominal value for V_{ds} .

A good EI means a 1D potential distribution in a device (as in the long-channel case), whereas poor EI means a 2D potential distribution that results in the 2D parasitic effects. A simple relationship between those two has been established, as follows:⁵

$$SCE \approx 2.0 \times \Phi_d \times EI$$

$$DIBL \approx 2.5 \times V_{ds} \times EI$$

where Φ_d is the source-to-channel junction built-in voltage, V_{ds} is the drain-to-source bias, and EI is given by:

$$EI \equiv \left(1 + \frac{X_j^2}{L_{el}^2} \right) \frac{T_{ox_el}}{L_{el}} \frac{T_{dep}}{L_{el}}$$

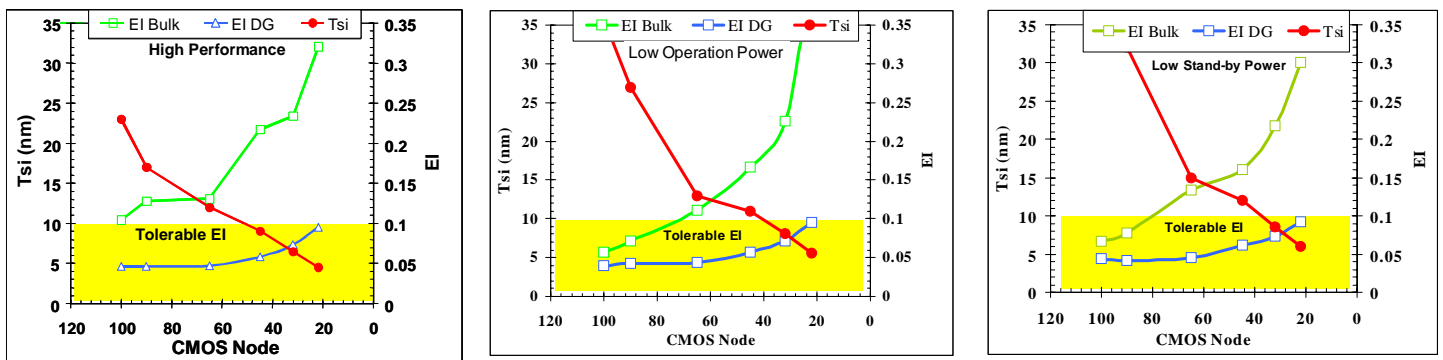
where X_j denotes the junction extension depth, L_{el} denotes the electrical channel length (junction-to-junction distance), T_{ox_el} denotes the effective electrical oxide thickness in inversion (equal to the sum of the equivalent oxide thickness of

⁵ T. Skomicki, invited talk, Proc. ESSDERC (September 2000), 19–33.

the gate dielectric, the gate polydepletion and the so-called “dark space”), and T_{dep} denotes the depletion depth in the channel. (“Dark space” is the distance the inversion charge layer peak is set back in the channel from the SiO_2/Si interface due to quantization of the energy levels in the channel quantum well.)

The strength of non-classical CMOS structures, in particular of UTB devices, is clearly shown by this expression when applying the translations of parameters relevant to UTB devices (refer to Table 60). Replacing X_j and T_{dep} by T_{si} (UTB single gate) or $T_{si}/2$ (UTB double gate) permits a considerable reduction in the X_j/L_{el} and T_{dep}/L_{el} ratios with the condition that silicon films of $T_{si} \ll X_j$, T_{dep} are available. The key question therefore is the extent to which body or channel thickness in advanced MOSFETs must be thinned to sustain EI.

Figure 39 compares the EI between bulk planar and double-gate devices throughout the span of nodes for the 2003 ITRS. It is encouraging to see that the T_{si} scaling, although very aggressive (4 nm and 5 nm Si films are required at the end of the roadmap for HP, and LOP/LSTP, respectively), has the potential to scale CMOS to the end of the roadmap with the SCE and DIBL at the same levels as the 90 nm node technologies.⁶ Note that the EI of planar bulk or classical devices is outside the allowed zone at the 100 nm node for HP, and near the 65 nm node for LOP and between the 90 nm and 65 nm nodes for LSTP products, respectively.



For double-gate devices the aggressive silicon film thickness scaling (down to 4 nm for high-performance devices and down to 5 nm for LOP and LSTP) ensures the EI to stay within the acceptable or tolerable range until the end of CMOS scaling.

Figure 39 Estimation of Electrostatic Integrity (EI) for Bulk and Double-gate FETs

Sustaining the $I_{on}-I_{off}$ Ratio—The technological maturity of some performance boosters is higher than that of others. For example strained-silicon channel devices already have been announced as being incorporated into the CMOS 65 nm node, whereas the metallic source/drain junction concept is in the research phase. Without attempting precise predictions on the introduction node for a given technology performance booster, the following sequence is suggested as a plausible scenario for their sequential introduction:

- Strained-Si channels
- UTB single-gate FETs
- Metallic-gate electrode
- UTB double-gate FETs
- Ballistic or quasi-ballistic transport
- Reduced fringing (and/or overlap) capacitance
- Metallic source/drain junction

⁶ $EI \leq 10\%$ (meaning DIBL of $<25\% V_{ds}$) is assumed as the acceptable range as represented as a yellow region in Figure 39.

MASTAR calculation with translation of technology boosters according to Table 60

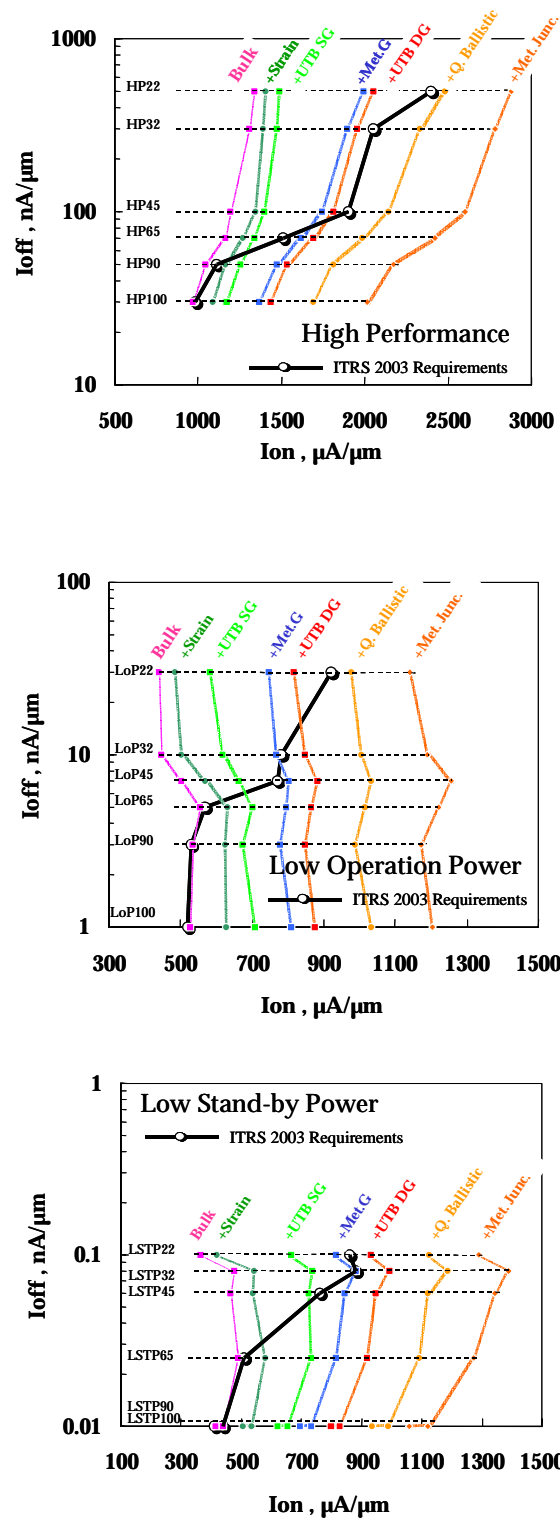


Figure 40 Impact of the Technology Boosters on HP, LOP, and LSTP CMOS Roadmaps in Terms of $I_{on}:I_{off}$ Ratio

Figure 40 shows the evolution of the $I_{\text{off}}-I_{\text{on}}$ Roadmaps (HP, LOP, and LSTP) due to introduction of the technology performance boosters as defined in Table 60, according to the above sequence and in a cumulative way. The planar bulk device is basically sufficient for satisfying the CMOS ($I_{\text{on}}-I_{\text{off}}$) specifications up to 90 nm node for HP and up to 65 nm node for LOP and LSTP. Beyond these nodes, introduction of technology performance boosters becomes mandatory for meeting the specifications. Exceeding the specifications appears possible if all boosters considered are co-integrated. It is also to be noted that the HP products use the greatest number of performance boosters (all except the metallic S/D junctions) to address the entire HP roadmap, whereas the LSTP roadmap can be satisfied with UTB single metallic gate devices.

The above analysis assumes that the I_{off} current is determined by the maximum allowed source/drain subthreshold leakage current (*refer to the PIDS logic technology requirements tables, note [5]*). The maximum gate leakage current is related to the maximum source/drain leakage current at threshold. For this to be true, high- κ dielectrics need to be introduced in 2006 for LOP and LSTP and in 2007 for high-performance logic.

Boosting the Speed (CV/I)—Certain performance boosters may lead to an increase in I_{on} at the same rate as the increase in C_{gate} , thus producing a small or negligible effect on CV/I (for example, see metallic gate in Table 60). Others, such as fringing or overlap capacitance, may reduce C_{gate} without altering I_{on} . The evolution of the intrinsic device speed $(CV/I)^{-1}$ as impacted by the performance boosters may thus be somewhat different than the evolution of the $I_{\text{on}}-I_{\text{off}}$. Figure 41 shows rough estimates for the evolution of the intrinsic device speed for the consecutive CMOS nodes. Up to the 65 nm node the optimized scaling strategy (basically equal to the ITRS 2001) is sufficient for the LOP and LSTP products to achieve an annual performance increase of 17%-per-year. HP products again require the most aggressive use of the performance boosters, such as requiring strained-Si channels beginning at the 65 nm node. Beyond this node, a sequential introduction of performance boosters is mandatory for maintaining the 17%-per-year performance improvement rate. At the 22 nm node, fringing (and/or overlap) capacitance needs to be reduced to meet the speed requirements of HP and LOP products. However, co-integrating the boosters up to and including the quasi-ballistic transport, according to the sequence presented in Table 60, can satisfy the requirements for LSTP. It is encouraging to see that the metallic junction booster is not employed within the current Roadmap, thus leaving a margin for its prolongation beyond the 22 nm node without any loss in the performance improvement rate.

MASTAR calculation with translation of technology boosters according to Table 60

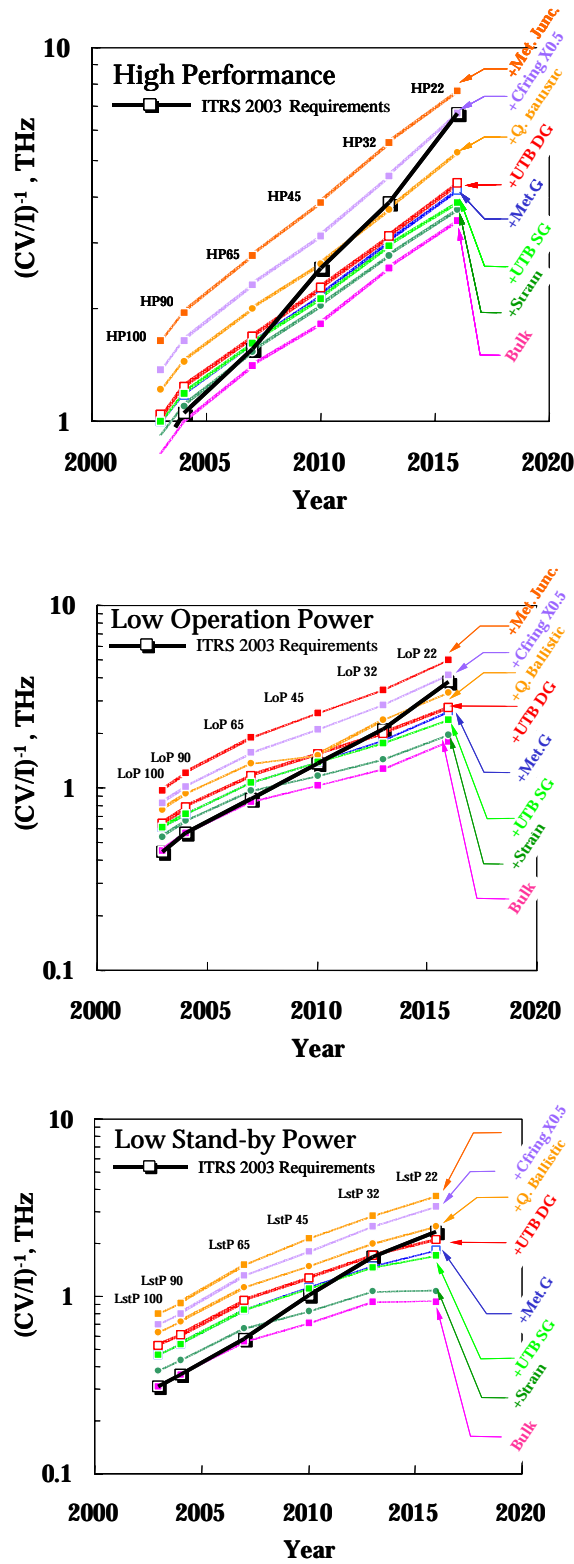


Figure 41 Impact of the Technology Boosters on HP, LOP, and LSTP CMOS Roadmaps in Terms of Device Intrinsic Speed ($f=1/(CV/I)$)

MEMORY DEVICES

INTRODUCTION

The memory technologies tabulated below are a representative sample of published 2003 research efforts selected to describe some attractive alternative approaches. Historically, very few memory research options yield practical memory devices, and including a particular approach here does not in any way constitute advocacy or endorsement. Conversely, not including a particular concept in this subsection does not in any way constitute rejection of that approach. This listing does point out that existing research efforts are exploring a variety of basic memory mechanisms. These mechanisms include charge isolated by surrounding dielectrics; charge held in place by Coulomb blockade potential; resistance change caused by chemical phenomena; and resistance change caused by material phase change. Table 61 is an organization or taxonomy of the emerging memory technologies into four categories. A strong theme is to merge each of these memory options into a CMOS technology platform in a seamless manner. Fabrication is viewed as modification of or addition to a CMOS platform technology. A goal is to present the end user with a device that looks like a familiar silicon memory chip. Because all of these approaches attempt to mimic and improve on the capabilities of present day memory technologies, parameters are provided for the current dominant volume produced memories of DRAM and Flash NOR technologies as benchmarks. Table 62a shows projected key parameters, which are estimates for potential performance of different emerging research memory devices at their maturity based on calculations and early experimental demonstrations. These parameters reflect a consensus of experts in this area. Table 62b contains up-to-date experimental values of these parameters reported in the cited technical references.

MEMORY TAXONOMY

Table 61 provides a simple way to categorize memory technologies. In this scheme, equivalent functional elements that make up a cell are identified. For example, the familiar DRAM cell that consists of an access transistor and a capacitor storage node is labeled as a 1T1C technology. Other technologies such as MRAM where data is stored as the spin state in a magnetic material can be represented as a 1T1R technology. Here the resistance “R” indicates that the cell readout is accomplished by sensing the current through the cell. The utility of this form of classification reflects the trend to simplify cells (i.e., reduce cell area) by reducing the number of equivalent elements to a minimum. Thus, early in the development of a given technology it is common to see multi-transistor multi-x (x equals capacitor or resistor) cells. As learning progresses, the structures are scaled down to a producible 1T1x form. The near ideal arrangement is to incorporate the data storage element directly into the transistor structure such that a 1T cell is achieved.

An important property that differentiates emerging technologies is whether data can be retained when power is not present. Non-volatile memory offers essential use advantages, and the degree to which non-volatility exists is measured in terms of the length of time that data can be expected to be retained. Volatile memories also have a characteristic retention time that can vary from milliseconds to (for practical purposes) the length of time that power remains on.

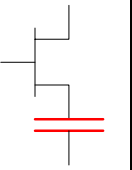
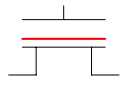
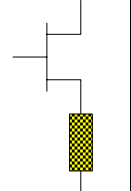
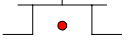
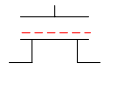
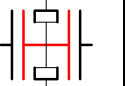
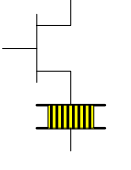
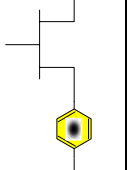
Table 61 Memory Taxonomy

<i>Cell Element</i>	<i>Type</i>	<i>Non-volatility</i>	<i>Retention Time</i>
1T1R	MRAM	Non-volatile	>10 years
	Phase Change Memory	Non-volatile	>10 years
	Polymer RAM	Non-volatile	>10 years
	Molecular memory	Volatile	>days
	Insulator Resistance Change Memory	Volatile	>years
1T1C	DRAM	Volatile	~seconds
	FeRAM	Non-volatile	>10 years
1T	FB DRAM	Volatile	<seconds
	Flash Memory	Non-volatile	>10 years
	SONOS	Non-volatile	>10 years
	Nano Floating Gate Memory	Non-volatile	>10 years
Multiple T	SRAM	Volatile	large
	STTM	Volatile	small
	Single Electron Memory	Volatile	large

*1T1R—1 transistor–1 resistor 1T1C—1 transistor–1 capacitor 1T—1 transistor FB DRAM—floating body DRAM
Multiple T—multiple transistor STTM—scaleable 2-transistor memory⁷*

⁷ J. H. Yi, W. S. Kim, S. Song, Y. Khang, H.J. Kim, J. H. Choi, H. H. Lim, N. I. Lee, K. Fujihara, H.K. Kang, J. T. Moon, and M. Y. Lee, "Scalable Two-transistor Memory (STTM)," *IEDM (2001)*, 36.1.1–36.1.4.

Table 62a Emerging Research Memory Devices—Projected Parameters

Storage Mechanism	Present Day Baseline Technologies		Phase Change Memory*	Floating Body DRAM	Nano-floating Gate Memory**	Single/Few Electron Memories**	Insulator Resistance Change Memory**	Molecular Memories**
								
Device Types	DRAM	NOR Flash	OUM	1TDRAM	Engineered tunnel barrier or nanocrystal	SET	MIM	Bi-stable switch
Availability	2004	2004	~2006	~2006	>2006	>2007	~2010	>2010
Cell Elements	1T1C	1T	1T1R	1T	1T	1T	1T1R	1T1R
Initial F	90 nm	90 nm	100 nm	70 nm	80 nm	65 nm	65 nm	45 nm
Cell Size	$8F^2$ $0.065 \mu m^2$	$12.5F^2$ $0.101 \mu m^2$	$\sim 6F^2$ $0.06 \mu m^2$	$\sim 4F^2$ ^[A] $0.0049 \mu m^2$	$\sim 6F^2$ $0.038 \mu m^2$	$\sim 6F^2$ $0.025 \mu m^2$	$\sim 6F^2$ $0.025 \mu m^2$	Not known
Access Time	<15 ns	~80 ns	<100 ns	<10 ns ^[A,B]	<10 ns	<10 ns	Slow	~10 ns
Store Time	<15 ns	~1 ms	<100 ns	<10 ns ^[A,B]	<10 ns	<100 ns	<100 ns	~10 ns
Retention Time	64 ms	10–20 yrs	>10 yrs	<10 ms ^[A]	>10 yrs	~100 sec	~1 year	~1 month
E/W Cycles	Infinite	1E5	>1E13	>1E15 ^[A]	>1E6	>1E9	>1E3	>1E15
General Advantages	<ul style="list-style-type: none"> Density Economy 	<ul style="list-style-type: none"> Non-volatile Multi-bit cells 	<ul style="list-style-type: none"> Non-volatile Low power Rad hard Multi-bit cells 	<ul style="list-style-type: none"> Density Economy 	<ul style="list-style-type: none"> Non-volatile Fast read and write Multi-bit cells 	<ul style="list-style-type: none"> Density Low power 	<ul style="list-style-type: none"> Low voltage Multi-bit cells 	<ul style="list-style-type: none"> Density Low power 3D potential Defect tolerant
Challenges	<ul style="list-style-type: none"> Scaling 	<ul style="list-style-type: none"> Scaling 	<ul style="list-style-type: none"> Large E/W current New materials and integration 	<ul style="list-style-type: none"> Need SOI Retention versus scaling Dopant fluctuation Endurance 	<ul style="list-style-type: none"> Material quality 	<ul style="list-style-type: none"> Dimension control for RT operation Background charge disturb 	<ul style="list-style-type: none"> New materials and integration Slow access Speed versus R trade-off 	<ul style="list-style-type: none"> Volatile Thermal stability
Maturity	Production	Production	Development	Demonstrated	Research	Research	Research	Research
Research Activity****			3***	3	61	40	3	43

Notes for Table 62a:

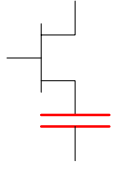
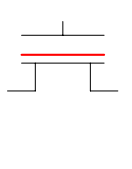
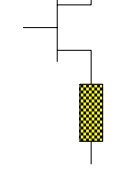
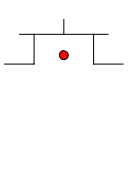
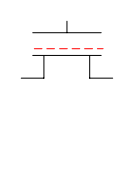
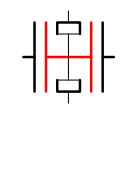
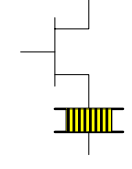
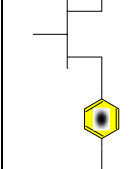
* Numerical data correspond to PCM parameters at the year of introduction and they reflect a consensus of the experts in the area based on experimental results (unpublished).

** Numerical data are estimates for potential performance of these memories based on calculations and early experimental demonstrations.

*** The basic research on phase change materials and their device applications was done in the 1960–70s. Currently, this technology is in development stage. There are only few publications on PCM in technical journals. Some information is available in conference literature, e.g., IEDM 2001, ISSCC 2002, VLSI 2003.

**** The number of referred articles in technical journals that appeared in the Science Citation Index database for 1/1/2001–6/4/2003.

Table 62b Emerging Research Memory Devices—Experimental Parameters

Storage Mechanism	Baseline 2004 Technologies		Phase Change Memory*	Floating Body DRAM	Nano-floating Gate Memory	Single/Few Electron Memories	Insulator Resistance Change Memory [C,D,E]	Molecular Memories
								
Device Types	DRAM	NOR Flash	OUM	1TDRAM	Engineered tunnel barrier or nanocrystal	SET	MIM	Bistable switch
Availability	2004	2004	~2006	~2006	>2006	>2007	~2010	>2010
Cell Elements	1T1C	1T	1T1R	1T	1T	1T	1T1R	1T1R
F Value	90 nm	90 nm	100 nm	130 nm ^[A,B]	80 nm	50 nm ^[G]	Not known	40–150 nm
Cell Size	8F ² 0.065 μm ² 1T	12.5F ² 0.101 μm ² 1T	~6F ² 0.06 μm ² 1T	9 to 13F ² ^[B]	4 to 10F ² 0.04 μm ²	200F ² ^[G] ~0.5 μm ²	80 μm ² ^[C]	9F ² ~0.01 μm ² ^[I]
Access Time	<15 ns	~80 ns	<100 ns	3 ns ^[A,B]	80 ns ^[F]	Not known	2 ms ^[C]	Not known
Store Time	<15 ns	~1 ms	<100 ns	3 ns ^[A,B]	100 ns ^[F]	5 ns ^[G]	100 ns ^[C]	~sec ^[I]
Retention Time	64 ms	10–20 yrs	>10 yrs	10–15 ms ^[B] (85°C)	>1 week ^[F]	>1 min ^[G]	1 year ^[D]	440 sec ^[H] ~month ^[I]
E/W Cycles	Infinite	>1E5	>1E13	Not known	1E9 ^[F]	Not known	>1E3 ^[D]	1E2 ^[I]
Advantages	See Table 62a							
Challenges								
Research Activity								

Note for Tables 62b:

* Numerical data correspond to PCM parameters at the year of introduction and they reflect a consensus of the experts in the area based on experimental results (unpublished).

References for Tables 62a and 62b:

[A] S. Okhonin, M. Nagoga, J.M. Sallese, and P. Fazan, "A Capacitor-less 1T-DRAM Cell," *IEEE Electron Dev. Lett.* 23, 2002, 85.

[B] P.C. Fazan, S. Okhonin, M. Nagoga, J.M. Sallese, "A Simple 1-Transistor Capacitor-less Memory Cell for High Performance Embedded DRAM," *IEEE 2002 Custom Integrated Circuits Conf.*

[C] A. Beck, J.G. Bednorz, C. Gerber, C. Rossel, and D. Widmer, "Reproducible Switching Effect in Thin Oxide Films for Memory Applications," *Appl. Phys. Lett.* 77, 2000, 139.

[D] Y. Watanabe, J.G. Bednorz, A. Bietsch, Ch. Gerber, D. Widmer, A. Beck, S. J. Wind, "Current-driven Insulator-conductor Transition and Non-volatile Memory in Chromium-doped SrTiO₃ Single Crystals," *Appl. Phys. Lett.* 78, 2001, 3738.

[E] C. Rossel, G.I. Meijer, D. Bremaud, D. Widmer, "Electrical Current Distribution Across a Metal-insulator-metal Structure During Bistable Switching," *J. Appl. Phys.* 90, 2001, 2892.

[F] S. Tiwari, et al., "A Silicon Nanocrystals Based Memory," *Appl. Phys. Lett.* 68, 1996, 1377.

[G] N.J. Stone, H. Ahmed, and K. Nakazato, "A High-speed Silicon Single-electron Random Access Memory," *IEEE Electron Dev. Lett.* 20, 1999, 583.

[H] C. Li, D. Zhang, X. Liu, S. Han, Tao Tang, C. Zhou, W. Fan, J. Koehne, Jie Han, M. Meyyappan, A.M. Rawlett, D.W. Price, J.M. Tour, "Fabrication Approach for Molecular Memory Arrays," *Appl. Phys. Lett.* 82, 2003, 645.

[I] Y. Chen, D.A.A. Ohlberg, X.M. Li, D.R. Stewart, R.S. Williams, J.O. Jeppesen, K.A. Nielsen, J.F. Stoddart, D.L. Olynick, E. Anderson, "Nanoscale Molecular-switch Devices Fabricated by Imprint Lithography," *Appl. Phys. Lett.* 82, 2003, 1610.

MEMORY DEVICES—DEFINITION AND DISCUSSION OF TABLE ENTRIES

Phase Change Memory—Phase change memory (PCM) also called Ovonic unified memory (OUM) is based on a rapid reversible phase change effect in some materials under influence of electric current pulses. The OUM uses the reversible structural phase-change in thin-film material (e.g., chalcogenides), which in turn changes the electrical resistivity of the material as the data storage mechanism. The small volume of active media acts as a programmable resistor between a high and low resistance with $>40\times$ dynamic range. 1's and 0's are represented by crystalline versus amorphous phase states of the active material. Phase states are programmed by the application of a current pulse through a MOSFET that drives the memory cell into a high or low resistance state, depending on current magnitude. Data is read by measuring resistance changes in the cell. PCM cells can be programmed to intermediate resistance values, such as for multi-state data storage. Since the energy required for phase transformation decreases with cell size, the write current scales with cell size, thus facilitating memory scaling. PCM devices have fast access time, long endurance, and good data retention. One of the challenges for PCM is to reduce the programming current to the level that is compatible with the minimum MOS transistor drive current for high-density integration. Currently, the programming current in the chalcogenide based PCM is substantially high. The lowest programming current reported at this point is 0.1–0.2 mA/device.⁸ This high current limits the minimum width of a MOS transistor needed to drive and sustain this current, which results in a larger cell size. A fully integrated PCM memory array, using a 0.24 μm MOSFET as the cell access transistor was recently reported.⁹ For this example, the minimum programming current is about 2.0 mA.

Floating body DRAM—Floating body DRAM is a 1-transistor capacitorless DRAM using the body charging of a partially-depleted SOI transistor to store the logic “1” or “0” states.^{10, 11, 12, 13, 14} The write operation is performed by using impact ionization to generate an excess charge in the floating body. The excess body charge alters the threshold voltage, and, thereby, the source/drain current of the transistor. Information is read by comparing the source/drain current of the selected cell to the current of a reference cell. This 1T cell may allow for very dense memory arrays, particularly for embedded applications. However, the retention time decreases with scaling, and for smaller devices, the retention time is too low for standalone DRAM applications.¹⁵ Also there are concerns about reliability and endurance issues due to memory operation that is based on energetic processes such as impact ionization. The primary application area for the floating body capacitorless memory is embedded DRAM.

Nanofloating gate memory (NFGM)—NFGM includes several possible evolutions of conventional floating gate memory. There are two major approaches proposed to improve performance of floating gate memory cells, as follows: 1) engineered tunnel barrier and 2) nano-sized memory node. Engineered tunnel barrier includes crested barrier floating gate memory¹⁶ and phase-state low-electron-number drive memory (PLED).¹⁷ The crested barrier concept uses a stack of insulating materials to create a special shape of barrier enabling effective Fowler-Nordheim tunneling into and out of the storage node.^{18, 19} In the PLED memory electrons are injected into the memory node through stacked multiple tunnel junctions with a double-gate structure. Engineered tunnel barriers serve to increase the read/write performance of memory cells while sustaining a long retention time typical for floating gate memories. The approach of NFGM with engineered tunnel barriers is currently in concept stage, since no memory operation has been experimentally demonstrated. In the

⁸ Y.H. Ha, et al., “An Edge Contact Type Cell for Phase Change RAM Featuring Very Low Power Consumption,” *VLSI Technology Symp.*, (2003), 175–176.

⁹ Y.N. Hwang, et al., “Full Integration and Reliability Evaluation of Phase-Change RAM Based on 0.24 μm -CMOS Technologies,” *VLSI Technology*, 2003, 173–174.

¹⁰ S. Okhonin, M. Nagoga, J.M. Sallese, and P. Fazan, “A Capacitor-less 1T-DRAM Cell,” *IEEE Electron Dev. Lett.* 23, 2002, 85.

¹¹ P.C. Fazan, S. Okhonin, M. Nagoga, J.M. Sallese, “A Simple 1-Transistor Capacitor-less Memory Cell for High Performance Embedded DRAM,” *IEEE 2002 Custom Integrated Circuits Conf.*

¹² T. Ohsawa, K. Fujita, T. Higashi, Y. Iwata, T. Kjiyama, Y. Asao, and K. Sunouchi, “Memory Design Using a One-transistor Gain Cell on SOI,” *IEEE J. Solid-State Circ.* 37, 2002, 1510.

¹³ C. Kuo, T.J. King, C. Hu, “A Capacitorless Double-gate DRAM Cell,” *IEEE Electron Dev. Lett.* 23, 2002, 345.

¹⁴ C. Kuo, T.J. King, C. Hu, “A Capacitorless Double-gate DRAM Cell Design for High Density Applications,” *IEDM* (2002).

¹⁵ P.C. Fazan, S. Okhonin, M. Nagoga, J-M Sallese, “A Simple 1-Transistor Capacitor-less Memory Cell for High Performance Embedded DRAM,” *IEEE 2002 Custom Integrated Circuits Conf.*

¹⁶ K.K. Likharev, “Layered Tunnel Barriers for Nonvolatile Memory Devices,” *Appl. Phys. Lett.* 73, 1998, 2137–2139.

¹⁷ K. Nakazato, K. Itoh, H. Mizuta, and H. Ahmed, “Silicon Stacked Tunnel Transistor for High-speed and High-density Random Access Memory Gain Cells,” *Electronics Lett.* 35, 1999, 848–850.

¹⁸ K.K. Likharev, “Layered Tunnel Barriers for Nonvolatile Memory Devices,” *Appl. Phys. Lett.* 73, 1998, 2137–2139.

¹⁹ H. Mizuta, H.O. Müller, K. Tsukagoshi, D. Williams, Z. Durrani, A. Irvine, G. Evans, S. Amakawa, K. Nakazato, and Haroon Ahmed, “Nanoscale Coulomb Blockade Memory and Logic Devices,” *Nanotechnology* 12, 2001, 155–159.

20 Emerging Research Devices

NFGM with nano-sized memory node, the floating gate consists of multiple²⁰ or single²¹ silicon nanocrystal dots. The multiple floating dots are separated and independent, and electrons are injected to the dots via different paths. The endurance and retention problem can be much improved in multidot (nanocrystal) memory. NFGM with nanosized memory node is sometimes referred as single electron memory.²² While most NFGM use standard MOSFETs for reading, some NFGM use semiconductor nanowires or carbon nanotubes. In one realization of Si nanowire memory, the channel of a polycrystalline Si nanowire FET is modulated with small charges trapped in localized areas naturally formed within the channel.^{23, 24} While this approach suffers lack of reproducibility, the elimination of the bulk MOSFET allows a significant reduction of the cell size. A 128 Mbit memory based on silicon nanowires has been demonstrated.²⁵ Another realization uses a semiconductor nanowire (Si, InP, GaP) FET functionalized with redox active molecules.²⁶ Recently, memories using carbon nanotubes were demonstrated.^{27, 28, 29, 30, 31} As this discussion shows, many concepts of NFGM with nanosized memory node, i.e., multidot, single dot, and nanowire memories have been experimentally demonstrated, but much work remains to prove their viability and high-yield manufacturability.

Single/few electron memory—In single electron devices electron movement (e.g., the addition or subtraction of an electron to a small 3D “island,” or quantum dot) is controlled with integer electron precision. Injection of each electron on to the quantum dot occurs through a tunneling barrier and is controlled by a separate gate electrode via the “Coulomb blockade” effect. In such quantum dots, electrons are confined electrostatically in all three dimensions, forming a small island of electrons, which is bounded on all sides by potential walls. This electron island can accommodate only an integer number of electrons, and these electrons can occupy only certain discrete energy states. Connected through tunneling barriers, the conductance of the dot exhibits strong oscillations as the voltage of a gate electrode is varied. Each successive conductance maximum corresponds to the discrete addition of a single electron to the dot. A basic component of single electron memory is the single electron transistor (SET). The SET is composed of a quantum dot connected to an electron source and to a separate electron sink through tunnel junctions with electron injection controlled by a gate electrode. Several concepts of single electron memory have been experimentally demonstrated,³² including a SET/FET hybrid.³³ Two major disadvantages of all single electron memories reported so far are very low operating temperature of 4.2–20K and sensitivity to background charges.

Insulator Resistance Change Memory—The basic component of this memory is a metal-insulator-metal (MIM) structure, using insulators, such as Cr-doped (Ba, Sr)TiO₃ or SrZrO₃,^{34, 35, 36} that show reproducible hysteresis in the leakage

²⁰ S. Tiwari, et al., “A Silicon Nanocrystals Based Memory,” *Appl. Phys. Lett.* 68, 1996, 1377.

²¹ X. Tang, X. Baie, J.P. Colinge, A. Crahay, B. Katschmarsyj, V. Scheuren, D. Spote, N. Reckinger, F. Van de Wiele, and V. Bayot, “Self-aligned Silicon-on-insulator Nano Flash Memory Device,” *Solid-State Electronics* 44, 2000, 2259–2264.

²² S.M. Sze, “Evolution of Nonvolatile Semiconductor Memory: from Floating-gate Concept to Single-electron Memory Cell,” in: *Future Trends in Microelectronics*, S. Luryi, J. Xu, and A. Zaslavsky, eds. John Wiley & Sons, Inc.: New York, NY, 1999, 291–303.

²³ K. Yano, T. Ishii, T. Hashimoto, F. Murai, and K. Seki, “Room-temperature Single-electron Memory,” *IEEE Trans. Electron. Dev.* 41, 1994, 1628–1638.

²⁴ K. Yano, T. Ishii, T. Sano, T. Mine, F. Murai, T. Hashimoto, T. Kobayashi, T. Kure, K. Seki, “Single-electron Memory for Giga-tera Bit Storage,” *Proc. IEEE* 87, 1999, 633–651.

²⁵ K. Yano, T. Ishii, T. Sano, T. Mine, F. Murai, T. Hashimoto, T. Kobayashi, T. Kure, K. Seki, “Single-electron Memory for Giga-tera Bit Storage,” *Proc. IEEE* 87, 1999, 633–651.

²⁶ X. Duan, Y. Huang, and C. M. Lieber, “Nonvolatile Memory and Programmable Logic from Molecule-gated Nanowires,” *Nano Letters* 2, 2002, 487.

²⁷ X. Duan, Y. Huang, and C. M. Lieber, “Nonvolatile Memory and Programmable Logic from Molecule-gated Nanowires,” *Nano Letters* 2, 2002, 487.

²⁸ W.B. Choi, S. Chae, E. Bae, and J.W. Lee, “Carbon-nanotube-based Non-volatile Memory with Oxide-nitride-oxide Film and Nanoscale Channel,” *Appl. Phys. Lett.* 82, 2003, 275.

²⁹ J.B. Cui, R. Sordan, M. Burghard, and K. Kern, “Carbon Nanotube Memory Devices of High Charge Storage Stability,” *Appl. Phys. Lett.* 81, 2002, 3260.

³⁰ M. Radosavljević, M. Freitag, K.V. Thadani, and A.T. Johnson, “Nonvolatile Molecular Memory Elements Based on Ambipolar Nanotube Field Effect Transistor,” *Nano Letters* 2, 2002, 761.

³¹ N. Yoneya, K. Tsukagoshi, Y. Aoyagi, “Charge Transfer Control by Gate Voltage in Crossed Nanotube Junction,” *Appl. Phys. Lett.* 81, 2002, 2250.

³² N.J. Stone, H. Ahmed, and K. Nakazato, “A High-speed Silicon Single-electron Random Access Memory,” *IEEE Electron Dev. Lett.* 20, 1999, 583.

³³ H. Mizuta, H.-O. Müller, K. Tsukagoshi, D. Williams, Z. Durrani, A. Irvine, G. Evans, S. Amakawa, K. Nakazoto, and Haroon Ahmed, “Nanoscale Coulomb Blockade Memory and Logic Devices,” *Nanotechnology* 12, 2001, 155–159.

³⁴ A. Beck, J.G. Bednorz, C. Gerber, C. Rossel, and D. Widmer, “Reproducible Switching Effect in Thin Oxide Films for Memory Applications,” *Appl. Phys. Lett.* 77, 2000, 139.

³⁵ Y. Watanabe, J.G. Bednorz, A. Bietsch, C. Gerber, D. Widmer, A. Beck, S.J. Wind, “Current-driven Insulator-conductor Transition and Non-volatile Memory in Chromium-doped SrTiO₃ Single Crystals,” *Appl. Phys. Lett.* 78, 2001, 3738.

current. The write operation is performed by applying different voltages to the MIM structure, which results in reversible switching between a low-resistance and a high-resistance state. Multilevel switching can be achieved in this structure. Data is read by measuring resistance of the MIM structures at the voltages lower than the write voltages (typically the read voltage is less than 0.5 V). The retention time of such MIM structure can be quite large—1 year retention was experimentally demonstrated. While stable and reproducible hysteresis was reported in MIM structures, a practical memory cell integrated with a sense transistor has not been demonstrated. An apparent drawback of this memory is large access time due to charging a high- κ MIM capacitor in parallel with the MIM resistor.

Molecular Memory—Molecular memory is a broad term encompassing different proposals for using individual molecules as building blocks of memory cells in which one bit of information can be stored in the space of an atom or a molecule. One experimentally demonstrated approach is based on rapid reversible change of effective conductance of a molecule attached between two electrodes controlled by applied voltage.^{37, 38, 39, 40} In this molecular memory data are stored by applying external voltage that cause the transition of the molecule into one of two possible conduction states. Data is read by measuring resistance changes in the molecular cell. There are also concepts for combining molecular components with current memory technology, such as DRAM⁴¹ and floating gate memory. In these cases the molecular element may act as a nanosized resonant tunnel diode or ultimately small memory node. It should be noted however, that the most recent work suggests that many of the earlier reported experimental results on electron transport through molecules were affected by experimental artifacts, such as formation of metal filaments along the molecule attached between two metal electrodes.⁴² Consequently, the knowledge base of molecular electronics needs further work.

LOGIC DEVICES

INTRODUCTION

The scaling of CMOS device and process technology, as it is known today, likely will end by the 16 nm node (7 nm physical channel length) by 2019. This end of scaling will be due to several concurrent fundamental and practical limits related to transistor operation and manufacturability. Fundamental limits include sustaining viable transistor operation and limiting thermal dissipation to manageable limits, both of which are common to all charge based logic devices and independent of device structure and material properties. The grand challenge, then, is to invent and develop one or more new technologies based on something other than electronic charge that will extend the scaling of information processing technologies through multiple generations beyond 2019. Undoubtedly, there will be opportunities for innovation and invention to extend CMOS devices and perhaps integrate them with other charge based logic devices in the near term beyond 2019, but these should be thought of as transitional technologies that will form a bridge to new highly scalable approaches.

Such new technologies must meet certain fundamental requirements and possess certain compelling attributes to justify the very substantial investments that will be necessary to build a new infrastructure. First and foremost, any new information processing technology must do the following:

1. Extend microelectronics orders of magnitude beyond the domain of CMOS and be capable of integration on or with a CMOS platform. This will require one or more of the following:
 - Functionally scaleable by several orders of magnitude beyond CMOS
 - High information/signal processing rate and throughput
 - Energy dissipation per functional operation substantially less than CMOS

³⁶ C. Rossel, G.I. Meijer, D. Bremaud, D. Widmer, "Electrical Current Distribution Across a Metal-insulator-metal Structure During Bistable Switching," *J. Appl. Phys.* 90, 2001, 2892.

³⁷ Y. Chen, D.A.A. Ohlberg, X.M. Li, D.R. Stewart, R.S. Williams, J.O. Jeppesen, K.A. Nielsen, J.F. Stoddart, D.L. Olynick, E. Anderson, "Nanoscale Molecular-switch Devices Fabricated by Imprint Lithography," *Appl. Phys. Lett.* 82, 2003, 1610.

³⁸ Y. Chen, G.Y. Jung, D.A.A. Ohlberg, X.M. Li, D.R. Stewart, J.O. Jeppesen, K.A. Nielsen, J.F. Stoddart, R.S. Williams, "Nanoscale Molecular-switch Crossbar Circuits," *Nanotechnology*, 14 (4), 2003, 462–468.

³⁹ Y. Luo, C.P. Collier, J.O. Jeppesen, K.A. Nielsen, E. Delonno, G. Ho, J. Perkins, H.R. Tseng, T. Yamamoto, J.F. Stoddart, J.R. Heath, "Two-dimensional Molecular Electronics Circuits," *ChemPhysChem* 3, 2002, 519.

⁴⁰ C. Li, D. Zhang, X. Liu, S. Han, T. Tang, C. Zhou, W. Fan, J. Koehne, J. Han, M. Meyyappan, A.M. Rawlett, D.W. Price, J.M. Tour, "Fabrication Approach for Molecular Memory Array," *Appl. Phys. Lett.* 82, 2003, 64.

⁴¹ J. Berg, S. Bengtsson, P. Lundgren, "Can Molecular Resonant Tunneling Diodes be Used for Local Refresh of DRAM Memory Dells?" *Solid-State Electron.* 44, 2000, 2247.

⁴² Robert F. Service, "Next-generation Technology Hits an Early Mid-life Crisis," *October 24, 2003, Science Vol. 302, 556–559.*

22 Emerging Research Devices

- Minimum scaleable cost per function
 - Room temperature operation
2. Provide a means for an energy restorative functional process to sustain steady state operation (e.g., in traditional devices provide a gain mechanism.)

The emerging logic technologies are tabulated in two separate but related tables to accurately portray both the current status and the long-term potential for the individual entries. Both tables contain the same set of rows that parameterize device operation. Table 63a contains entries that characterize the ultimate performance limit projected in the open literature for each device. In general, the values of these parameters are based on dimensional arguments and known physical limits. Table 63b contains the experimental performance recently reported for the same devices and parameters as those presented in Table 63a. A comparison of the two tables gives an indication of the current state of development of each technology.

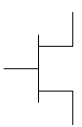
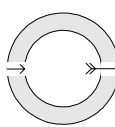
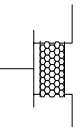
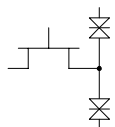
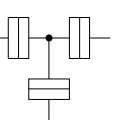
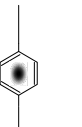
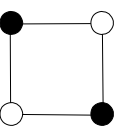
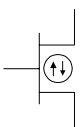
The last row in Table 63a contains the number of papers published in the last three years on the particular device technology. It is meant to be a gauge of the amount of research activity currently taking place in the research community and it is a primary metric that determines which of the candidate devices are included in these tables. The rows labeled “Advantages” and “Challenges” list the characteristics associated with a particular device where there is a high degree of consensus. The tables have been extensively footnoted and details may be found in the indicated references. The text associated with the tables gives a brief summary of the operating principles of each device and as well as significant issues that were not captured in the tables.

As with memory concepts, including a particular concept or approach in this subsection does not in any way constitute approval, advocacy, or endorsement. Conversely, not including a particular concept in this subsection does not in any way constitute rejection of that approach. Additional information on these devices is contained in an excellent summary of nanoelectronic devices entitled *Technology Roadmap for Nanoelectronics*, produced by the European Commission’s IST program (Future and Emerging Technologies).⁴³ The interested reader is also referred to an excellent, comprehensive, and authoritative book published in 2003, entitled *Nanoelectronics and Information Technology*.⁴⁴

⁴³*Technology Roadmap for Nanoelectronics, 2nd Edition, R. Compano, ed., np, November 2000.*

⁴⁴*Nanoelectronics and Information Technology, Rainer Waser, ed. Wiley-VCH, 2003.*

Table 63a Emerging Research Logic Devices—Projected Parameters

Availability Sequence	1	2	2-3	2-3	4	5	6	
Device								
	FET	RSFQ ^[A,B,C]	1D structures	Resonant Tunneling Devices	SET	Molecular	QCA ^[D]	Spin transistor
Types	<ul style="list-style-type: none"> Si CMOS 	<ul style="list-style-type: none"> JJ 	<ul style="list-style-type: none"> CNT FET NW FET NW hetero-structures Crossbar nanostructure 	<ul style="list-style-type: none"> RTD-FET RTT 	<ul style="list-style-type: none"> SET 	<ul style="list-style-type: none"> 2-terminal 3-terminal FET 3-terminal bipolar transistor NEMS Molecular QCA 	<ul style="list-style-type: none"> E: QCA** M: QCA** 	<ul style="list-style-type: none"> Spin FET (SFET) Spin-valve transistor (SVT)
Supported Architectures	<ul style="list-style-type: none"> Conventional 	<ul style="list-style-type: none"> Pulse 	<ul style="list-style-type: none"> Conventional Cross-bar 	<ul style="list-style-type: none"> Conventional CNN 	<ul style="list-style-type: none"> CNN 	<ul style="list-style-type: none"> Memory-based QCA 	<ul style="list-style-type: none"> QCA 	<ul style="list-style-type: none"> Quantum Programmable logic
Cell Size (spatial pitch)	100 nm*	0.3 μm	100 nm*	100 nm*	40 nm	Not known	60 nm	100 nm*
Density (device/cm ²)	3E9	1E6	3E9	3E9	6E10	1E12	3E10	3E9
Switch Speed	700 GHz	1.2 THz	Not known	1 THz	1 GHz	Not known	30 MHz	700 GHz
Circuit Speed	30 GHz	250–800 GHz	30 GHz	30 GHz	1 GHz	<1 MHz (NEMS)	1 MHz	30 GHz
Switching Energy, J***	2×10 ⁻¹⁸	2×10 ⁻¹⁹ (Nb) [>1.4×10 ⁻¹⁷]	2×10 ⁻¹⁸	>2×10 ⁻¹⁸	1×10 ⁻¹⁸ [>1.5×10 ⁻¹⁷] ^[C]	1.3×10 ⁻¹⁶ (NEMS)	[E:>1×10 ⁻¹⁸] ^[E] M:>4×10 ⁻¹⁷	2×10 ⁻¹⁸
Binary Throughput, GBit/ns/cm ²	86	0.4	86	86	10	N/A	0.06	86
Gain	Must be >>1 for all devices. See Table 63b for experimental values							
Operational Temperature	RT	<ul style="list-style-type: none"> 4 K (Nb) 77 K (HTS) 20 K (MgB₂) 	RT	RT	20 K	RT	E:QCA Cryogenic M:QCA RT	<ul style="list-style-type: none"> Cryogenic (SFET) RT (SVT)
CD Tolerance	Critical	Not critical	Not critical	Very critical	Very critical	Not critical	Very critical <2% (M: QCA)	Critical
Materials System	Si	Nb HTS	CNT Si III-V	III-V Si-Ge	III-V Si	C-60	Al/Al ₂ O ₃ (E: QCA)	<ul style="list-style-type: none"> III-V (SFET) Si/FM (SVT)
Most Complex Circuit Demonstrated	See Table 63b							

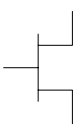
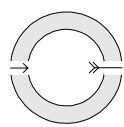
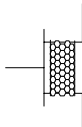
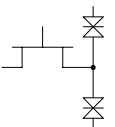
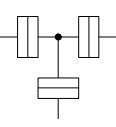
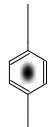
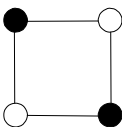
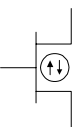
*Scaling of these structures is the same as the scaling of FETs

**E:QCA or E:—electric QCA M:QCA or M:—magnetic QCA

***The value in the [] is the value that includes cooling energy. If an ideal Carnot refrigerator is used for cooling to the operation temperature T_c , the total switching energy $E_{sw} > E_c \cdot \frac{300}{T_c}$, where E_c is the net switching energy, when cooling energy is not taken into account.

24 Emerging Research Devices

Table 63a Emerging Research Logic Devices—Projected Performance at Maturity (continued)

Availability Sequence	1	2	2-3	2-3	4	5	6	
Device								
	FET	RSFQ ^[A,B,C]	1D structures	Resonant Tunneling Devices	SET	Molecular	QCA ^[E]	Spin transistor
Advantages		<ul style="list-style-type: none"> Very high circuit speed 		<ul style="list-style-type: none"> Density (smaller cell size) 		<ul style="list-style-type: none"> Identity of individual switches on sub-nm level Potential solution to interconnect problem 	<ul style="list-style-type: none"> Morphological simplicity 	
Challenges		<ul style="list-style-type: none"> Cryogenic operations 		<ul style="list-style-type: none"> Stand-by power Process integration 	<ul style="list-style-type: none"> Cryogenic operations 			<ul style="list-style-type: none"> Low spin injection efficiency Short coherence time
Research Activity****		80	103	71	53	62	35	46

**** The number of referred articles in technical journals that appeared in the Science Citation Index database for 1/1/2001–7/8/2003.

References for Table 63a:

[A] Kadin, et al., "Can RSFQ Logic Circuits be Scaled to Deep Submicron Junctions?" *IEEE Trans. Appl. Supercond.* 11 (2001), 1050.

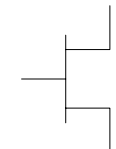
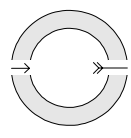
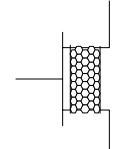
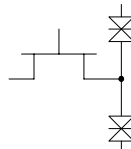
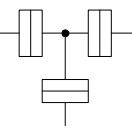
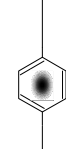
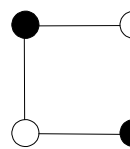
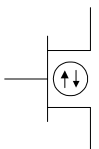
[B] Y. Naveh, D.V. Averin, and K.K. Likharev, "Physics of High Jc Nb/AlOx/Nb Josephson Junctions and Prospects for their Applications," *IEEE Trans. Appl. Supercond.* 11 (2001), 1056.

[C] A.W. Kleinasser, "High Performance Nb Josephson Devices for Petaflops Computing," *IEEE Trans. Appl. Supercond.* 11 (2001), 1043.

[D] A.O. Orlov, et al., "Experimental Demonstration of Clocked Single-electron Switching in Quantum-dot Cellular Automata," *Appl. Phys. Lett.* 77 (2000), 295.

[E] J. Timler and C.S. Lent, "Power Gain and Dissipation in Quantum-dot Cellular Automata," *J. Appl. Phys.* 91 (2002), 823. The energy/bit represents a conservative estimate, corresponding to an irreversible abrupt switching cell, which dissipates the full value of its kink energy every clock cycle (see also Fig. 11 and its caption in this reference.)

Table 63b Emerging Research Logic Devices—Experimental Parameters

Availability Sequence		1	2	2-3	2-3	4	5	6
Device								
	FET	RSFQ ^[Z]	1D Structures	Resonant Tunneling Devices	SET	Molecular	QCA	Spin transistor
Types		JJ	CNT FET ^[F,G,H]	MOBILE ^[I] MVL RTT ^[J]	AND ^[L] NOT ^[M,N] OR ^[O]	2-terminal [V,W,X,Y]	E: QCA ^{[P,Q,R]*} M: QCA ^[S,T]	Spin-valve ^[U]
Supported Architectures	See Table 63a							
Cell Size (Spatial Pitch)	100 nm	46 μm	10 μm	3 μm ^[K]	10 μm ^[I] 100 μm ^[L,M] 1 μm ^[N]	120 nm	E: 5.8 μm ^{[P]*} M: 250 nm ^[S,T]	2 mm
Density (Devices/cm ²)	3E9	5E4	Not known	Not known	Not known	6E9	M: 2E9	Not known
Switch Speed	700 GHz	51–80 GHz	220 Hz	700 GHz	1 MHz ^[L]	2 Hz	Not known	Not known
Circuit Speed	30 GHz	20 GHz	Not known	Not known	Not known	Not known	E: 0.03–0.1 Hz ^[Q,R,U] M: 27 Hz ^[T]	Not known
Switching Energy, J**	2×10 ⁻¹⁸	1.14×10 ⁻¹⁷ [>8×10 ⁻¹⁶]	10 ⁻¹⁰	10 ⁻¹³ [Z]	8×10 ⁻¹⁷ [I] [>1.3×10 ⁻¹⁴]	10 ⁻⁹	E: 4×10 ⁻²³ [>8×10 ⁻¹⁹] [R] M: 6×10 ⁻¹⁸	Not known
Binary Throughput, Gbit/ns/cm ²	86	9E-4	Not known	Not known	Not known	6.4E-9	M: 4E-8	Not known
Gain			g _m =3 μA/V	β=50 ^[J]	1.3–2 ^[I]	Not known	E: 2.07 ^[U]	α=10 ⁻³
Operational Temperature	RT	4 K	RT	RT	1.8 K ^[L] 27 K ^[M,N] 1.5 K ^[O]	RT	E: 15–70 Mk ^[Q,R,U] M: RT ^[S,T]	RT
CD Tolerance	See Table 63a							
Materials System	Si	Nb/AlO _x /Nb	CNT	III-V Si-Ge Si	GaAs ^[L,N] Si ^[L,M]	Organic molecules	Al/Al ₂ O ₃ (E: QCA) Permalloy/Si (M: QCA)	GaAs/Fe/Au
Most Complex Circuit Demonstrated		8-bit microprocessor	Ring oscillator ^[G,H]	ADC NAND/NOT 16-bit TSRAM	2 bit adder ^[M]	64-bit cross-bar array ^[W]	Latch (E: QCA) Shift register (M: QCA)	Not known
Advantages	See Table 63a							
Challenges								
Research Activity								

*E: QCA or E:—electric QCA M: QCA or M:—magnetic QCA

**The value in the [] is the value that includes cooling energy. If an ideal Carnot refrigerator is used for cooling to the operation temperature T_c, the total switching energy E_{sw} > E_c · $\frac{300}{T_c}$, where E_c is the net switching energy, when cooling energy is not taken into account.

26 Emerging Research Devices

References for Tables 63a and 63b:

- [A] Kadin, et al., "Can RSFQ Logic Circuits be Scaled to Deep Submicron Junctions?" *IEEE Trans. Appl. Supercond.* 11, 2001, 1050.
- [B] Y. Naveh, D.V. Averin, and K.K. Likharev, "Physics of High Jc Nb/AlOx/Nb Josephson Junctions and Prospects for their Applications," *IEEE Trans. Appl. Supercond.* 11, 2001, 1056.
- [C] A.W. Kleinasser, "High-performance Nb Josephson Devices for Pedaflops Computing," *IEEE Trans. Appl. Supercond.* 11, 2001, 1043.
- [D] A.O. Orlov, et al., "Experimental Demonstration of Clocked Single-electron Switching in Quantum-dot Cellular Automata," *Appl. Phys. Lett.* 77, 2000, 295.
- [E] J. Timler and C.S. Lent, "Power Gain and Dissipation in Quantum-dot Cellular Automata," *J. Appl. Phys.* 91, 2002, 823. The energy/bit represents a conservative estimate, corresponding to an irreversible abrupt switching cell, which dissipates the full value of its kink energy every clock cycle (See also Fig. 11 and its caption in this reference).
- [F] S.J. Wind, J. Appenzeller, R. Martel, V. Derycke, P. Avouris, "Vertical Scaling of Carbon Nanotube Field-effect Transistors Using Top Gate Electrodes," *Appl. Phys. Lett.* 80, 2002, 3817–3819.
- [G] A. Javey, Q. Wang, A. Ural, Y.M. Li, H.J. Dai, "Carbon Nanotube Transistor Arrays for Multistage Complementary Logic and Ring Oscillators," *Nano Lett.* 2, 2002, 929–932.
- [H] A. Bachtold, P. Hadley, T. Nakanishi, C. Dekker, "Logic Circuits with Carbon Nanotube Transistors," *Science* 294, 2001, 1317–1319.
- [I] U. Auer, et al., "Low-voltage Mobile Logic Module Based on Si/SiGe Interband Tunneling Devices," *IEEE Electron Dev. Lett.* 22, 2001, 215.
- [J] M.A. Reed, et al., "Realization of a Three-terminal Resonant Tunneling Device: The Bipolar Quantum Resonant Tunneling Transistor," *Appl. Phys. Lett.* 54, 1989, 1034.
- [K] P. Fau, et al., "Fabrication of Monolithically-integrated InAlAs/InGaAs/InP HEMTs and InAs/AlSb/GaSb Resonant Interband Tunneling Diodes," *IEEE Trans. Electron Dev.* 48, 2001, 1282.
- [L] K. Tsukagoshi, et al., "Operation of Logic Function in a Coulomb Blockade Device," *Appl. Phys. Lett.* 73, 1998, 2515.
- [M] Y. Ono, et al., "Si complementary Single-electron Inverter with Voltage Gain," *Appl. Phys. Lett.* 76, 2000, 3121.
- [N] Y. Ono, et al., "Fabrication of Single-electron Transistor and Circuits Using SOIs," *Solid-State Electron.* 46, 2002, 1723.
- [O] S. Kasai and H. Hasegawa, "GaAs and InGaAs Single Electron Hexagonal Nanowire Circuits Based on Binary Decision Diagram Logic Architecture," *Physica E* 13, 2002, 925.
- [P] A.O. Orlov, et al., "Experimental Demonstration of Clocked Single-electron Switching in Quantum-dot Cellular Automata," *Appl. Phys. Lett.* 77, 2000, 295.
- [Q] A.O. Orlov, et al., "Experimental Demonstration of a Latch in Clocked Quantum-dot Cellular Automata," *Appl. Phys. Lett.* 78, 2001, 1625.
- [R] R.K. Kummamuru, et al., "Power Gain in a Quantum-dot Cellular Automata Latch," *Appl. Phys. Lett.* 81, 2002, 1332.
- [S] R.P. Cowburn and M.E. Welland, "Room Temperature Magnetic Quantum Cellular Automata," *Science* 287, 2000, 1466.
- [T] D.A. Allwood, et al., "Submicrometer Ferromagnetic NOT Gate and Shift Register," *Science* 286, 2002, 2003.
- [U] R. Sato and K. Mizushima, "Spin-valve Transistor with an Fe/Au/Fe(001) base," *Appl. Phys. Lett.* 79, 2001.
- [V] Y. Chen, D.A.A. Ohlberg, X.M. Li, D.R. Stewart, R.S. Williams, J.O. Jeppesen, K.A. Nielsen, J.F. Stoddart, D.L. Olynick, E. Anderson, "Nanoscale Molecular-switch Devices Fabricated by Imprint Lithography," *Appl. Phys. Lett.* 82, 2003, 1610.
- [W] Y. Chen, G.Y. Jung, D.A.A. Ohlberg, X.M. Li, D.R. Stewart, J.O. Jeppesen, K.A. Nielsen, J.F. Stoddart, R.S. Williams, "Nanoscale Molecular-Switch Crossbar Circuits," *Nanotechnology*, 14 (4), 2003, 462–468.
- [X] M.A. Reed, et al., "Molecular Random Access Memory Cell," *Appl. Phys. Lett.* 78, 2001, 3735.
- [Y] Y. Luo, C.P. Collier, J.O. Jeppesen, K.A. Nielsen, E. Delonno, G. Ho, J. Perkins, H.R. Tseng, T. Yamamoto, J.F. Stoddart, J.R. Heath, "Two-dimensional Molecular Electronics Circuits," *ChemPhysChem* 3, 2002, 519.
- [Z] M. Dorojevets, "An 8-bit FLUX-1 RSFQ microprocessor built in 1.75- μ m Technology," *Physica C* 378, 2002, 1446.

LOGIC DEVICES—DEFINITION AND DISCUSSION OF TABLE ENTRIES

Resonant tunnel devices^{45, 46}—The resonant tunnel devices for logic applications include resonant tunnel transistors (RTT) and hybrid devices incorporating resonant tunneling diodes and one or more FETs (RTD-FET). The RTDs are two terminal devices that intrinsically have a very high switching speed and exhibit a region of negative differential resistance in their I-V curves. These two characteristics make them potentially attractive as high speed switching devices. If two RTDs are connected in series, they have two stable operating points and can switch between the two stable points very quickly if a third terminal is added that can act as a gate. However, since the peak current through an RTD depends exponentially on the barrier thickness, it is inherently difficult to get reproducible device operation unless the gate also controls the peak current. Controlling the peak current is usually done by integrating a transistor with the double RTD on a common substrate.⁴⁷ This approach results in complex, epitaxially grown structures requiring very good control of film thicknesses.

Integration of a transistor with a pair of RTDs introduces additional delays to the inherently fast bistable switching times associated with capacitive charging and discharging of the transistor gate stack. The operational speed of the integrated device can be an order of magnitude slower than the intrinsic switching speeds of the RTDs themselves. Additional challenges include a limited dynamic range of 10 rather than the factor of 10^5 that CMOS digital circuit designers require and the inherent complexity of the integrated structure, which limits the dimensional scaling of the devices. The complexity of the hybrid devices makes them large with experimental spatial pitch values of order $3\ \mu\text{m}$ being reported. Another issue is fabrication of silicon or silicon-germanium tunnel diodes with peak-to-valley ratios $> 4\text{--}5$ required to obtain stable circuit operation.

Adding a control terminal to RTDs extends their usability to a variety of applications. This approach has been used to build resonant tunneling transistors (RTT).⁴⁸ RTTs have a negative transconductance that can be used in several logic circuits, e.g., in XOR gate with only one transistor.⁴⁹

Overall, the resonant tunneling devices may be useful for certain niche applications requiring high speed and low dynamic range, provided the manufacturing issues associated with uniformity of the tunneling barrier can be resolved.

Single-electron transistors (SETs)—SETs⁵⁰ are three-terminal switching devices that can transfer electrons from source to drain one by one. The principle of operation of the single-electron transistor proposed for logic applications is the same as that of the single-electron memory element described in the Memory section. The structure of SETs is almost the same as that of FETs. The important difference, however, is that in a SET the channel is separated from source and drain by tunneling junctions, and the role of channel is played by a quantum dot. Operational parameters of SETs depend on the size of the quantum dot. Operation of SET circuits is generally limited to very low temperatures. Estimates^{51, 52} of the logic gate parameters, based on 2 nm SETs, are maximum operation temperature $T \sim 20\ \text{K}$, integration density $n \sim 10^{11}\ \text{cm}^{-2}$, and speed of the order of 1 GHz.

Implementation of SETs in logic circuits is thought to increase packing density and possibly to decrease power consumption. There are two approaches for implementing logic operations in the circuits of single-electron devices. The first approach is to represent one bit by a single electron and use a SET to transfer electrons one by one. In the second approach, each bit is represented by more than one electron, while a SET is used to switch the current on/off.

⁴⁵ D. J. Paul, B. Coonan, G. Redmond, G. M. Crean, B. Holländer, S. Mantl, I. Zozoulenko, K.F. Berggren, "Silicon Quantum Integrated Circuits," in: *Future Trends in Microelectronics*, eds., S. Luryi, J. Xu, and A. Zaslavsky. John Wiley & Sons, Inc.: New York, NY, 1999. 183–192.

⁴⁶ *Nanoelectronics and Information Technology*. Rainer Wasser, ed. Wiley-VCH, 2003. 416–424.

⁴⁷ P. Fau, et al., "Fabrication of Monolithically-integrated InAlAs/InGaAs/InP HEMTs and InAs/AlSb/GaSb Resonant Interband Tunneling Diodes," *IEEE Trans. Electron Dev.* 48, 2001, 1282.

⁴⁸ M.A. Reed, et al., "Realization of a Three-terminal Resonant Tunneling Device: the Bipolar Quantum Resonant Tunneling Transistor," *Appl. Phys. Lett.* 54, 1989, 1034.

⁴⁹ P. Fau, et al., "Fabrication of Monolithically-integrated InAlAs/InGaAs/InP HEMTs and InAs/AlSb/GaSb resonant Interband Tunneling Diodes," *IEEE Trans. Electron Dev.* 48, 2001, 1282.

⁵⁰ *Nanoelectronics and Information Technology*. Rainer Wasser, ed. Wiley-VCH, 2003. 425–444.

⁵¹ R.H. Chen, A.N. Korotkov, and K.K. Likharev, "Single-electron Transistor Logic," *Appl. Phys. Lett.* 68, 14, 1994.

⁵² C.P. Gerousis and S. M. Goodnick, "Simulation of Single-Electron Tunneling Circuits," *Phys. Stat. Sol. B* 233, 2002, 113.

28 Emerging Research Devices

Two general disadvantages of SET logic circuits are low error immunity and limited fan-out. The low error immunity is due to very high probability of false bit occurrence, for example, as a result of thermal noise or background charge. The limited fan-out, an intrinsic feature of SETs, is the inability for a given SET gate to drive more than one other gate.

Low error tolerance and low fan-out make it difficult for SETs to compete with CMOS logic. Correspondingly it is important to develop an operation scheme, under which the functionality of SET circuits can surpass that of conventional CMOS circuits. A programmable SET logic and multi-value logic are examples of possible functionality improvement by utilization of SETs.⁵³

*Rapid Single Flux Quantum (RSFQ)*⁵⁴—RSFQ logic is a dynamic logic based upon a superconducting quantum effect, in which the storage and transmission of flux quanta (Fluxon) defines the device operation. The basic RSFQ structure is a superconducting ring that contains one Josephson Junction (JJ) plus an external resistive shunt. The storage element is the superconducting inductive ring and the switching element is the Josephson Junction. RSFQ dynamic logic uses the presence or absence of the flux quanta in the closed superconducting inductive loop to represent a bit as a “1” or “0,” respectively. The circuit operates by temporarily opening the Josephson Junction, thereby ejecting the stored flux quanta. A quantized voltage pulse is then generated across the Josephson Junction. This voltage pulse is propagated down a superconducting transmission line and can be used to trigger other RSFQ structures in various combinations to form complex circuits. As this quantum effect occurs at a macroscopic scale, sub-micron lithography is not a prerequisite. With RSFQ, circuit speeds above 100 GHz, perhaps up to 800 GHz, are possible. RSFQ circuits are currently built on low temperature superconducting Josephson Junctions (~5 K). However, high temperature superconductors may eventually be exploited. RSFQ devices need extreme cooling because the device operating temperature is lower than the critical temperature of the bulk superconductor material. The availability of adequate cooling systems, which comply with needed specifications (temperature, size, weight, dimensions, etc.) in the limits of reasonable prices, is one of the most important drawbacks for the market introduction of this technology.

Recent studies⁵⁵ have addressed the ultimate scalability of RSFQ circuits and have shown analytically that it should be possible to scale RSFQ circuits to 0.3 μm and a frequency of 250 GHz. Additional dimensional scaling will increase device density but not result in increased performance. This limit is based on the fact that the junction will become self-shunting at 0.3 μm and further scaling of the circuit will not reduce the resistance or the effective time constant. A second limit to scaling of RSFQs and their component JJs is a limit imposed by thermal dissipation. The most important requirement for the operation of the RSFQ is for it to remain below the critical temperature of the superconductor. The average RSFQ switch operation dissipates approximately 50 nW and at a device pitch of 0.3 μm , this corresponds to a system-cooling requirement of 50 W/cm². A third limit to scaling of current, commercially available RSFQ devices (which use niobium/aluminum oxide/niobium trilayer junctions) relates to their relatively low critical current density j_c of ~1 kA/cm². A direct consequence of the low j_c is that the area of the layout required for the shunt resistor is quite large compared to the junction dimensions themselves, thereby limiting the maximum device density. Recent studies⁵⁶ suggest that using JJs with higher critical current density of ~200 kA/cm² can increase the device density by a factor of 1000.

The principle advantage associated with RSFQ circuits is their very high operational speeds of up to 770 GHz in simple flip-flop circuits. However, their ultimate scaling density appears to be limited due to the factors discussed above which also limits their binary information throughput, as defined above, to be much less than that for scaled silicon. There also appears to be no viable way to avoid cryogenic operation, which imposes a substantial cost burden. Commercial application to niche applications where speed is the dominant requirement is likely to continue but wide scale application is unlikely.

Quantum cellular automata (QCA)—In the QCA paradigm a regular array of cells, each interacting with its neighbors, is employed in a locally interconnected architecture. The coupling between the cells is given by their electrostatic interactions and not by wires. Such arrays are in principle capable of encoding digital information.

⁵³ R.H. Chen, A.N. Korotkov, and K.K. Likharev, “Single-electron Transistor Logic,” *Appl. Phys. Lett.* 68, 14, 1954.

⁵⁴ K. Block, K. Track and M. Rowell, “Superconducting ICs: the 100 GHz Second Generation,” *IEEE Spectrum*, December 2000, 40–46.

⁵⁵ Kadin, et al., “Can RSFQ Logic Circuits be Scaled to Deep Sub-micron Junctions?” *IEEE Transactions on Applied Superconductivity*, Vol 11, No. 1, March, 2001.

⁵⁶ Y. Naveh, D.V. Averin, and K.K. Likharev, “Physics of High J_c Nb/AlOx/Nb Josephson Junctions and Prospects for their Applications,” *IEEE Trans. Appl. Supercond.* 11, 2001, 1056.

The archetype QCA is the electronic QCA (E: QCA).^{57, 58} A single E: QCA cell is made up of 4, 5, or 6 individual dots or isolated conductive islands. In a 4-dot cell, the quantum dots occupy the corners of a square cell. Due to electrostatic repulsion, the charges will occupy the dots in diagonally opposite corners of the cell and form two bistable states representing +1 and -1. The physical mechanisms of interaction between the dots are the Coulomb interaction and quantum mechanical tunneling. If the cells are arranged in a regular square grid then long-established cellular automata theory can be applied, together with its extension, cellular non-linear (or neural) network (CNN) theory to describe the information processing algorithm. Use of CNN allows a large body of theory to be applied directly to QCA architectures, which are further described in the Architectures section.

Standard solid state QCA cell design considers the inter-dot distance in a cell of approximately 20 nm and the inter-cell distance of 60 nm.⁵⁹ A hypothetical single-molecule implementation of QCA cell has been proposed, which requires a molecule in which charge is localized on specific sites and can tunnel between those sites.⁶⁰ For this molecular QCA, the inter-dot distance is expected to be about 2 nm, and the inter-cell distance about 6 nm.

Any QCA functional logic unit consists of more than one QCA cell. For example, a NOT gate (inverter) consists of eleven QCA cells.⁶¹ Correspondingly, by assessing the potential of QCA for logic devices, it is important to consider the whole unit, not a single cell. Favorable assumptions, which can be found in the literature, suggest the intrinsic switching time for an individual QCA cell to be in the THz range.⁶² On the other hand, comparative analysis of circuit performance of QCA and CMOS against a representative computer task, suggests that realistic circuits of solid state QCA will have the maximum operating frequency of several MHz.^{63, 64} Small circuits of hypothetical molecular QCA might have the maximum operating frequency of several GHz, however, as the circuit size increases, capacitive loading effects will reduce the speed.

A severe problem of QCA arises from the fact that QCA are single electron devices, and correspondingly they are very sensitive to the background charge. Today, no viable solutions to the background charge immune single-electron systems are known.

Another serious drawback of QCA devices is that room temperature operation is not achievable with solid-state QCA systems. For the standard solid state QCA cell, the maximum operating temperature was estimated to about 7 K.⁶⁵ Molecular implementation seems to be the only possibility of fabricating large-scale QCA circuits operating at room temperature. The electronic QCA will suffer both from effectively low packing density and low operation speeds in comparison to CMOS if conventional designs and a 2D architecture are used.

In addition to electronic QCA, the concept of magnetic QCA (M: QCA), using ferromagnetic dots was proposed. Estimated minimum size of magnetic dots is about 20 nm, and the minimum size of magnetic QCA cells is about 100 nm. Optimistic estimates of switching energy and speed of M: QCA yield correspondingly 10^{-17} J/switch and 200 MHz.^{66, 67}

1D structures—Reduced or one-dimensional device structures include several device concepts, each containing a 1D structure (e.g., nanotube or nanowire) as a critical element. Several 1D devices have been demonstrated, including *carbon*

⁵⁷ C.S. Lent and P.D. Tougaw, "Dynamics of Quantum Cellular Automata," *J. Appl. Phys.* 80, 1996, 4722-4736

⁵⁸ C.S. Lent and P.D. Porod, "A Device Architecture for Computing with Quantum Dots," *Proc. IEEE* 85, 1997, 541-557.

⁵⁹ C.S. Lent and P.D. Tougaw, "Dynamics of Quantum Cellular Automata," *J. Appl. Phys.* 80, 1996, 4722-4736

⁶⁰ C.S. Lent, B. Isaksen, and M. Lieberman, "Molecular Quantum-dot Cellular Automata," *J. Am. Chem. Soc.* 125, 2003, 1056-1063.

⁶¹ C.S. Lent and P.D. Tougaw, "Dynamics of Quantum Cellular Automata," *J. Appl. Phys.* 80, 1996, 4722-4736

⁶² C.S. Lent and P.D. Tougaw, "Dynamics of Quantum Cellular Automata," *J. Appl. Phys.* 80, 1996, 4722-4736

⁶³ K. Nikolic, D. Berzon, M. Forshaw, "Relative Performance of Three Nanoscale Devices—CMOS, RTDs and QCAs—Against a Standard Computing Task," *Nanotechnology* 12, 2001, 38-43.

⁶⁴ L. Bonci, G. Iannaccone, and M. Macucci, "Performance Assessment of Adiabatic Quantum Cellular Automata," *J. Appl. Phys.* 89, 2001, 6435-6443.

⁶⁵ C.S. Lent and P.D. Tougaw, "Dynamics of Quantum Cellular Automata," *J. Appl. Phys.* 80, 1996, 4722-4736

⁶⁶ R.P. Cowburn and M.E. Welland, "Room Temperature Magnetic Quantum Cellular Automata," *Science* 287, 2000, 1466.

⁶⁷ D.A. Allwood, et al., "Submicrometer Ferromagnetic NOT Gate and Shift Register," *Science* 286, 2002, 2003.

30 Emerging Research Devices

nanotube (CNT) FETs,⁶⁸ *semiconductor nanowire (NW) FETs*,⁶⁹ *semiconductor nanowire heterostructures*,⁷⁰ and *crossbar nanostructures*.⁷¹

One-dimensional material systems offer two potential advantages over bulk material systems, thus making them the subject of intense research activity. The potential advantages most frequently cited are enhanced mobility relative to bulk systems and phase-coherent transport of electron wavefunctions. Enhanced mobility may lead to faster transistors and devices. Phase-coherent transport of a single electron wavefunctions may lead to new wave-dependent functionality in which the 1D nanowire behaves as an electron waveguide. This latter phenomenon could lead to new wave interference devices, similar in concept to RF waveguide and single-mode optical fiber structures. Wave interference device concepts proposed include the Y-branch switch,⁷² quantum interference transistor, employing the Aharonov-Bohm effect,⁷³ a directional coupler⁷⁴ and a stub transistor.⁷⁵ The 1D systems being studied include single crystal nanowire devices of Si and Ge, as well as carbon nanotube structures. The diameter of the 1D structures is 1–30 nm where, for the smaller diameters, room temperature quantum confinement can be significant.

Arguments vary regarding the theoretical potential of enhanced mobility in NWs. One argument⁷⁶ proposes that enhanced mobility is due to a reduced probability of electron-phonon scattering associated with the reduced density of states. Another⁷⁷ points out that quantum confinement increases the energy in the allowed phonon modes so the quantitative probability of a 180° scattering event may not be reduced. The experimental evidence on this point is unclear at the moment. In most experimental work, the measured mobility in NWs is very low. On the other hand, a most recent result was reported showing very high mobility in p-type B-doped Si NW 10–20 nm in diameter—the mobility in some NW samples was as high as 1350 cm²/V-s.⁷⁸

Carbon nanotubes are an important subset of 1D structures because their semiconducting band structure can vary from metallic to semiconducting to insulating. A carbon nanotube is a molecular “tube” or cylinder formed from an atomic “sheet” of carbon atoms. These carbon atoms are bonded together into an array of hexagons, which form a planar sheet, similar to an atomic sheet of graphite (resembling a planar assembly of open hexagons). This graphite-like (graphene) sheet is rolled up to form a carbon nanotube. Carbon nanotubes can have diameters between 1–20 nm and lengths from 100 nm to several microns. The tube diameter and just how the sheet of carbon hexagons is rolled up determine whether a tube is a semiconductor or a metal. If a tube is a semiconductor, the details of rolling also determine the energy bandgap and, therefore, the electronic properties of the tube. These bandgap energies range all the way from zero (like a metal) to values as large as silicon (1.1 eV), with many values in between. The tubes can be doped both p- and n-type making possible p-n junctions. Several groups have demonstrated p-FET device structures in which a gate electrode modulates the tunneling probability or conductivity of a source/channel Schottky tunnel barrier by a factor of 10⁵ or more, providing an I_{on}/I_{off} ratio similar to those for silicon MOSFETs.⁷⁹ Different simple circuits with CNT transistors were demonstrated,

⁶⁸ P.G. Collins and P. Avouris, “Nanotubes for Electronics,” *Scientific American*, December 2000, 62–69.

⁶⁹ Y. Cui, Z. Zhong, D. Wang, W.U. Wang, and C.M. Lieber, “High Performance Silicon Nanowire Field Effect Transistors,” *Nano Letters* 3 (2), 2003, 149–152.

⁷⁰ L.J. Lauhon, M.S. Gudiksen, C.L. Wang, C.M. Lieber, “Epitaxial Core-shell and Core-multishell Nanowire Heterostructures,” *Nature*, Vol. 420, 2002, 57–61.

⁷¹ T. Rueckes, K. Kim, E. Joselevich, G.Y. Tseng, C.L. Cheung, C.M. Lieber, “Carbon Nanotube-based Nonvolatile Random Access Memory for Molecular Computing,” *Science* 289, 2000, 94–97.

⁷² A.M. Song, M. Missous, P. Omling, A.R. Peaker, L. Samuelson, W. Seifert, “Unidirectional Electron Flow in a Nanometer-scale Semiconductor Channel: A Self-switching Device,” *Appl. Phys. Lett.* 83, 2003, 1881–1883.

⁷³ E.K. Heller, F.C. Jain, “Simulation of One-dimensional Ring Quantum Interference Transistors Using the Time-dependent Finite-difference Beam Propagation Method,” *J. Appl. Phys.* 87, 2000, 8080–8087.

⁷⁴ N. Tsukada, M. Gotoda, M. Nunoshita, T. Nishino, “Nonlinear Electron-wave Directional Coupler,” *Phys. Rev. B* 53, 1996, R7603–R7606.

⁷⁵ J. Appenzeller, C. Schroer, “Multimode Transport in a T-shaped Quantum Transistor,” *J. Appl. Phys.* 87, 2000, 3165–3167.

⁷⁶ P.L. McEuen, M.S. Fuhrer, H.K. Park, “Single-walled Carbon Nanotube Electronics,” *IEEE T Nanotechnol* 1, 2002, 78–85.

⁷⁷ S. Bandyopadhyay, A. Svizhenko, M. A. Stroscio, “Why Would Anyone Want to Build a Narrow Channel (Quantum Wire) Transistor?” *Superlat. and Microstr.* 27, 2000, 67–76.

⁷⁸ Y. Cui, Z. Zhong, D. Wang, W. U. Wang, and C. M. Lieber, “High Performance Silicon Nanowire Field Effect Transistors,” *Nano Letters* 3 (2), 2003, 149–152.

⁷⁹ S.J. Wind, J. Appenzeller, R. Martel, V. Derycke, P. Avouris, “Vertical Scaling of Carbon Nanotube Field-effect Transistors Using Top Gate Electrodes,” *Appl. Phys. Lett.* 80, 2002, 3817–3819.

such as NOT, NOR logic gates, a flip-flop, and a ring oscillator.^{80, 81} A complimentary voltage inverter (both the p- and n-channel FETs in this inverter were fabricated in a single carbon nanotube) has been demonstrated.⁸²

Similar to the CNT FET, a NW FET consists of one of two structures. The first is a device with a channel made out of a semiconductor NW having a diameter of 10–20 nm. The second structure consists of the cross intersection of two different NWs consisting of a p-Si channel and an n-GaN gate separated by a thin SiO_x dielectric.⁸³ These NW FETs have exhibited I_{on}/I_{off} ratios exceeding 10⁴ or 10⁵, respectively.

Even though the conductivity per unit width of a nanotube or a nanowire can be quite large, the fact that the transverse dimensions of a NT or NW FET are quantized in units of an individual tube diameter (1–20 nm) will cause the overall drive current produced by a single NT or NW device to be quite limited unless many 1D structures can be combined in parallel. Experimental attempts to do this have not been successful because the individual tubes have widely varying properties. Another problem associated with any 1D structure is the contact resistance between the 1D structure and the bulk material. This resistance has a minimum value associated with the quantum of resistance (12 KΩ) so even though transport in the body of a 1D structure may be ballistic, the transport through the structure will be limited by the electrical contacts.

Molecular devices—The concept of electronic components based on electronic transport through individual nanoscale molecules derives from the observation that molecules can be stable, molecules can be chemically self-assembled, and some molecules show bistable operation with respect electron transport. The principal problems associated with the use of molecules as components in nanoelectronic devices stem from the difficulties in attaching electrical connections to molecules.

Molecular electronic logic devices are assumed to be based on electron transport properties through a single molecule.⁸⁴ A potential applied to the molecule results in reconfiguration of the molecular components, or moieties, and a change in the molecule's electrical conduction properties.⁸⁵ The exact mechanism of charge transport in molecules is not well understood. One suggested model is the change in a molecule's electrical properties is caused by a change in the overlap of the orbitals in the molecule. The correct overlap of the molecular orbitals allows electrons to flow through the molecule. But when this overlap of orbitals is further changed (because the molecule has been twisted or its geometry has been otherwise changed by externally applied fields) the flow of electrons is impeded. It should be noted however, that the most recent work suggests that many of the earlier reported experimental results on electron transport through molecules were affected by experimental artifacts, such as formation of metal filaments along the molecule attached between two metal electrodes. Consequently, the knowledge base of molecular electronics needs further work.⁸⁶

Experimental demonstrations to date have been performed using both two-terminal devices^{87, 88, 89, 90} and three-terminal devices (without gain).⁹¹ Two terminal molecular devices currently being explored consist of thousands of molecules operating in parallel, e.g., as digital switches or as analog diodes. Other two terminal applications involve placing a single molecule between the intersection points of a crossbar array.⁹² In this configuration, two rectangular grids are laid

⁸⁰ A. Javey, Q. Wang, A. Ural, Y. M. Li, H. J. Dai, "Carbon Nanotube Transistor Arrays for Multistage Complementary Logic and Ring Oscillators," *Nano Lett.* 2, 2002, 929–932.

⁸¹ A. Bachtold, P. Hadley, T. Nakanishi, C. Dekker, "Logic Circuits with Carbon Nanotube Transistors," *Science* 294, 2001, 1317–1319.

⁸² V. Derycke, R. Martel, J. Appenzeller, and P. Avouris, "Carbon Nanotube Inter- and Intramolecular Logic Gates," *Nano Letters*, 1, 2001, 453.

⁸³ Y. Huang, X.F. Duan, Y. Cui, C.M. Lieber, "Gallium Nitride Nanowire Nanodevices," *Nano Lett.* 2, 2002, 101–104.

⁸⁴ J.C. Ellenbogen and J.C. Love, "Architectures for Molecular Electronic Computers: I. Logic Structures and an Adder Designed from Molecular Electronic Diodes," *Proc. IEEE* 88, 2000, 386–425.

⁸⁵ Y. Wada, "Prospects for Single Molecule Information Processing Devices," *Proc. IEEE* 89, 2001, 1147–1171.

⁸⁶ Robert. F. Service, "Next-generation Technology Hits an Early Mid-life Crisis," *October* 24, 2003, *Science* Vol 302, 556–559.

⁸⁷ Y. Chen, D.A.A. Ohlberg, X.M. Li, D.R. Stewart, R.S. Williams, J.O. Jeppesen, K.A. Nielsen, J.F. Stoddart, D.L. Olynick, E. Anderson, "Nanoscale Molecular-switch Devices Fabricated by Imprint Lithography," *Appl. Phys. Lett* 82, 2003, 1610.

⁸⁸ Chen Y., Jung G.Y., Ohlberg D.A.A., Li X.M., Stewart D.R., Jeppesen J.O., Nielsen K.A., Stoddart J.F., Williams R.S., "Nanoscale Molecular-switch Crossbar Circuits," *Nanotechnology*, 14 (4), 2003, 462–468.

⁸⁹ M.A. Reed, J. Chen, A.M. Rawlett, et al., "Molecular Random Access Memory Cell," *Appl. Phys. Lett* 78, 2001, 3735.

⁹⁰ Y. Luo, C.P. Collier, J.O Jeppesen, K.A. Nielsen, E. Delonno, G. Ho, J. Perkins, H.R. Tseng, T. Yamamoto, J.F. Stoddart, J.R. Heath, "Two-dimensional Molecular Electronics Circuits," *ChemPhysChem* 3, 2002, 519.

⁹¹ N.B. Zhitenev, A. Erbe, H. Meng, Z. Bao, "Gated Molecular Devices Using Self-assembled Monolayers," *Nanotechnology* 14, 2003, 254–257.

⁹² M.A. Reed, J. Chen, A.M. Rawlett, et al., "Molecular Random Access Memory Cell," *Appl. Phys. Lett* 78, 2001, 3735.

32 Emerging Research Devices

down perpendicular to one another and molecules are used to connect the points at which the two grids intersect. Individual molecules can then be addressed by using one grid as read lines and the other grid as write lines. By using molecules that display a region of negative differential resistance, it is possible to perform logic operations with such an array.

Another approach uses a C-60 molecule positioned between a source and drain on a thin, insulating substrate.⁹³ A back gate underneath the substrate controls the internal state of the C-60 molecule and the entire device has the same functionality as a FET with very high series resistance. The problems associated with contacts and contact resistance is very evident in this and other molecular logic devices.

Spin transistors—The spin transistor is restricted to mean three-terminal devices, namely spin-FET and spin-valve transistors, that modulate current through spin coupling effects. The spin-FET (proposed concept) is based on a narrow gap semiconductor FET with ferromagnetic source and drain contacts. The ferromagnetic source and drain enable injection and collection of spin-polarized electrons that are controlled by the gate voltage. The spin-valve transistor resembles a bipolar transistor. It consists of a semiconductor emitter and collector separated by a base region made out of a thin multi-layer ferromagnetic-nonmagnetic metal sandwich. The spin-valve transistors were experimentally demonstrated, however the spin-polarized current transfer ratio is very low. Spin-based devices discussed in this section do not include the more general category of “spintronic” devices such as spin-LEDs, spin-detectors, spin RTDs, optical switches, encoders, decoders, and modulators. The more general field of spintronics has been reviewed recently.⁹⁴

Another review⁹⁵ focuses on the spin-transistor and discusses three types of spin controlled current modulation devices. One original concept of a spin valve uses momentum dependent spin charge coupling to modulate the electron current into the collector. At the present time, no one has demonstrated this concept. Both the original spin-valve concept and its variations depend on the same microscopic interaction between spin orientations of individual particles and a macroscopic polarization of a ferromagnetic material that give rise to the giant magneto resistive effect. The performance of such devices is likely to be limited by the low dynamic range and energy efficiencies because they involve electron transport across potential barriers.

In contrast to the microscopic interactions underlying the spin-FET and spin-valve concepts, a spin-dependent macroscopic interaction has been proposed.⁹⁶ This approach utilizes the electromotive force produced by any non-equilibrium charge carrier population inversion. In this case, the energy-dependent Zeeman splitting of electron spins in a magnetic field will produce an electromotive force that can cause current flow across a heterojunction and can demonstrate current gain. The physical effect has been demonstrated and devices based on the effect have been proposed. The simplest configuration is two magnetically doped p/n junctions arranged back to back to form a bipolar transistor structure. In principle however, multiple junctions can be concatenated to obtain better dynamic range similar to the way read/write heads are constructed using sandwiches of giant magnetic ratio (GMR) materials.

The prevalence and enormous economic impact of spin-based transport on magnetic storage media continues to drive the search for similar devices that can be applied to logic technology. So far, no viable devices have been demonstrated or proposed but even the restricted field of spin-transistors is the subject of a great deal of research activity.

⁹³ H. Park, J. Park, A.K.L. Lim, E.H. Anderson, A.P. Alivisatos, P.L. McEuen, “Nanomechanical Oscillations in a Single-C-60 Transistor,” *Nature* 407, 2000, 57–60.

⁹⁴ S. Wolf, et al., “Spintronics: A Spin Based Electronics Vision for the Future,” *Science*, 294, Nov. 16, 2001.

⁹⁵ Das Sarma, et al., “Spin Electronics and Spin Computations,” *Solid State Communications*, 119, 2001, 207–215.

⁹⁶ Das Sarma, Fabian, and Zutic, “Spin Polarized Transport and its Applications,” *Archive of Condensed Matter*, Vol 1, June 2002.

EMERGING RESEARCH ARCHITECTURES

INTRODUCTION⁹⁷

This section describes the coupling of future nanoscale devices to new applications and the architectures needed to support them. *The definitions of architecture used in this section are discussed in the supplemental material.* Table 64 summarizes these architectural approaches.

The characteristics of nanoscale devices and fabrication methods that must be considered in developing appropriate circuits and computing architectures⁹⁸ include regularity of layout, unreliable device performance, device transfer functions, interconnect limitations, and thermal power generation. The regular layout is a result of the self-assembly methods that must be used at dimensions below those for which the standard “top-down” processing techniques are used. The device performance is a consequence of both the physical principles and the inherent variability associated with the nanoscale where it is estimated that percent quantities of devices will not function adequately for useful circuits. Device transfer functions include the need for gain so that complex circuits can be designed and input/output relationships can be realized that are useful for circuit design. Interconnect limitations come from the following two origins: 1) the geometrical challenge of accessing extremely small devices with connections that will transfer information at the needed speed and bandwidth, and 2) the transformation of interconnect dimensions from the nanoscale to the physical world of realizable system connections. Thermal power generation comes from the device switching energy and also the energy needed to drive signals through circuits. For electron transport devices such as MOS transistors, the switching energy is projected to be 2×10^{-18} J/switching transition at the 22 nm node, which will limit the useable combination of device density and speed.⁹⁹

The limitations of nanoscale devices impose restrictions on organizations that are available for future architectures. Local computing tiles composed of simple device structures have been proposed that are interconnected with nearest neighbors through crossbar interconnect arrangements that bound the devices. Such organizations satisfy constraints on device gain and interconnect parasitics. Other organizations are based on molecular devices inspired by biological systems with much larger circuit fan-out than used in today’s technology. Such circuits work by using chemical regulatory approaches. For all nanoscale organizations, the management of defective devices will be a critical element of any future architecture since the defect rates are expected to be much higher than current practice.

ARCHITECTURES—DEFINITION AND DISCUSSION OF TABLE ENTRIES

FINE-GRAINED PARALLEL IMPLEMENTATIONS IN NANOSCALE CELLULAR ARRAYS

For nanoscale devices, the integration level will be terascale (10^{12} devices/cm²). For this large number of devices, many new information processing and computing capabilities are possible in principle that would not be considered at the gigascale level of integration. For many reasons, these devices will need to be interconnected mostly locally and patterned in grids or arrays of cells. Devices such as quantum dots interconnected in regular arrays by local Coulomb charge interactions are being considered for terascale densities. Two architecture implementations proposed for these cellular arrays are QCA and CNN. Actually, QCA implementations can be regarded as a subset of CNNs,¹⁰⁰ but since they evolved separately, they are discussed separately. These implementations are particularly useful for hybrid analog/digital systems with data structures that map well to parallel processing.

Quantum Cellular Automata Architecture Implementations—The QCA paradigm is one in which a regular array of cells, each interacting with its neighbors, is employed in a locally interconnected manner. Such cells are typically envisioned to be electrostatically coupled quantum dots, or magnetic-field-coupled nanomagnets. Ongoing research is exploring QCA in various molecular structures as well.¹⁰¹ Therefore, there are no wires in the signal paths. If QCA cells are arranged in a

⁹⁷ The following discussion applies primarily to nanoscale devices that are dominated by quantum transport phenomena. Coherent quantum devices and architectures based on energy transition probabilities and phases are still in the research phase and, although mentioned briefly herein, a detailed discussion will be deferred until later revisions of the ITRS.

⁹⁸ Refer to the Supplemental Backup Section for a definition of architecture.

⁹⁹ V.V. Zhirnov, R.K. Cavin, J.A. Hutchby, G.I. Bourianoff, “Limits to Binary Logic Switch Scaling—A Gedanken Model,” *Proc. IEEE*, Vol. 91, 2003, 1934–1939.

¹⁰⁰ W. Porod, C.S. Lent, G. Toth, A. Csurgay, Y.F. Huang, R.W. Liu, n.t., *IEEE Abstracts*, 1997, 745.

¹⁰¹ C.S. Lent, n.t., *Science*, 288, 2000, 1597.

34 Emerging Research Devices

closely packed grid, then long-established cellular automata theory can be used to implement specific information processing algorithms. Also, QCA can be extended to cellular nonlinear (neural) networks discussed below. Thus a large body of theoretical algorithm implementations can be applied to QCA arrays. By departing from close-packed, regular grid structures, it is possible to use QCAs to carry out general logic functions and universal computing with modest efficiency. In addition to non-uniform layouts, QCAs need a spatially non-uniform “adiabatic clocking field” that controls cell switching from one state to another and allows them to evolve rapidly to a stable end state. The clock also produces some gain, non-linearity, and isolation between neighboring parts of a circuit. It is possible to construct a complete set of Boolean logic gates with QCA cells and to design arbitrary computing structures. The energy per switching transition, adjusted for the required cooling energy, is expected to be of order $3 \times 10^{-19} \text{J}$ to $3 \times 10^{-15} \text{J}^{102}$ at 100 GHz (as compared to values of $4 \times 10^{-18} \text{J}$ projected for CMOS at the 22 nm node). Although there will be no interconnect capacitance associated with these structures, there will be a significant capacitance associated with the inter-dot size and spacing geometries. Power gain in QCAs has been demonstrated by using energy from the clock.¹⁰³

Cellular Nonlinear Networks—A CNN is an array of mainly identical dynamical systems called cells that satisfy two properties as follows: 1) most interactions are local, within a distance of one cell dimension, and 2) the state variables are continuous valued signals (not digital). A template specifies the interaction between each cell and all its neighbor cells in terms of their input, state, and output variables. The interaction between the variables of one cell may be either a linear or nonlinear function of the variables associated with its neighbor cells. A cloning function determines how the template varies spatially across the grid and determines the dynamical response of the array to boundary values and initial conditions. Since the range of interaction and the connection complexity of each cell are independent of the number of cells, the architecture is extremely scalable, reliable, and robust. Programming the array consists of specifying the dynamics of a single cell, the connection template, and the cloning function of the templates. This approach is simpler than traditional VLSI design methodology since the functional components are simple and reusable.

CNNs can be used to implement Boolean logic as well as more complex functions such as majority gates, MUX gates, and switches. CNNs can simulate many mathematical problems such as diffusion and convection and nervous system functions. The CNN organization also lends itself to implementing defect management techniques as discussed below. Devices that can be used include quantum dots QCAs,^{104,105} SETs, and RTDs. Tunneling phase logic has been combined with CNN to enable neural-like spike switching waveforms and ultra-low power dissipation.¹⁰⁶

One caution concerning CNNs is that despite the potential applications discussed above, the only published application to date has been for analog image processing. However algorithms for pattern recognition and analysis can be implemented very efficiently in CNNs.

DEFECT TOLERANT ARCHITECTURE IMPLEMENTATIONS¹⁰⁷

The goal of fault and defect-tolerant implementations is to enable reliable circuits and computing using unreliable devices. Defects can occur as permanent defects from hardware manufacturing and as transient defects such as random charges that affect single-electron transistors. Defective devices may be functional but still not meet the tolerance and reliability requirements for effective large-scale circuit operation. These effects are expected to be particularly acute for quantum-dominated devices at the nano- and molecular scale and will require significant resources to control.¹⁰⁸

It is expected that the invention of nanoscale devices could eventually permit extremely large scales of integration, of the order 10^{12} devices per chip. However, it is almost certain that it will be very difficult to make nanoscale circuits with any degree of functional certainty. Furthermore, it is likely that the proposed nanoelectronic devices will be more fragile than conventional FETs and will be sensitive to external influences. Hence, fault-tolerant architectures will certainly be

¹⁰² J. Timler and C.S. Lent, “Power Gain and Dissipation in Quantum-Dot Cellular Automata,” *J. Appl. Phys.*, 91, 2002, 823.

¹⁰³ J. Timler and C.S. Lent, “Power Gain and Dissipation in Quantum-Dot Cellular Automata,” *J. Appl. Phys.*, 91, 2002, 823.

¹⁰⁴ G. Toth, C.S. Lent, P.D. Tougaw, Y. Brazhnik, W.W. Weng, W. Porod, R.W. Liu, and Y.F. Huang, “Quantum Cellular Neural Networks, Superlattices and Microstructures,” 20(4), 1996, 473–478.

¹⁰⁵ A.I. Csurgay, “Signal Processing with Near Neighbor Coupled Time Varying Quantum Dot Arrays,” *IEEE Trans. Circuits and Systems, -I: Fundamental Theory and Applications*, 47, 2000, 1212

¹⁰⁶ T. Yang, R. A. Kiehl, and L. O. Chua, “Tunneling Phase Logic Cellular Nonlinear Networks,” *International Journal of Bifurcation and Chaos in Applied Sciences and Engineering*, Vol. 11, 2001, 2895–2911.

¹⁰⁷ While defect tolerance in itself is not a separate class of architecture, its pervasive application to any architecture implemented using nanoscale device technologies is best treated by discussion in this separate section.

¹⁰⁸ J.R. Heath, et al., “A Defect Tolerant Computer Architecture: Opportunities for Nanotechnology,” *Science*, 280, 1998, 1716.

necessary in order to produce reliable systems that are immune to manufacturing defects and to transient errors such as noise, crosstalk, power-supply fluctuations, cosmic rays, and temperature or stray-charge variations.

Several techniques exist for overcoming the effects of inoperative devices. All of these techniques use the concept of redundancy in resources or in time. The most representative techniques are as follows: R -fold modular redundancy (RMR),¹⁰⁹ NAND multiplexing (NAND-M),¹¹⁰ and reconfiguration (RCF).¹¹¹ The effectiveness of RCF was successfully demonstrated on a massively parallel computer “Teramac.”¹¹² An analysis¹¹³ of fault tolerance of nanocomputers has recently been presented. Two characteristic parameters of a defect or fault-tolerant architecture are the amount of redundancy R and the allowable failure rate per device p_f . In this context, redundancy usually means static redundancy—redundant rows and columns, for example. Dynamic redundancy is used to catch and correct problems “on the fly” and is a more expensive use of resources. It is not clear how much dynamic redundancy will be needed at the nano and molecular levels until new computing models are developed.

The choice of defect- or fault-tolerant schemes may be both manufacturing and application specific. For example, although the RMR technique is the least effective, with the level of redundancy of $R=5$, one can achieve the same level of chip reliability, but with devices which are four orders of magnitude less reliable. The price for this improvement is that the effective number of devices is reduced to $N/5$ (and the p_f for each device must be smaller than 10^{-9} for $N=10^{12}$ devices). On the other hand, the reconfigurable computer can in principle handle extremely large manufacturing defect rates—in the limit, even approaching unity—but only at the expense of colossal amounts of redundancy. If one wishes to fabricate a chip containing the equivalent of many present-day workstations, then the device failure rate during manufacturing must be smaller than 10^{-5} . This may be difficult to achieve for nanoscale devices. RMR and NAND-M in general are not as effective as reconfiguration. However, if the dead devices cannot be located during manufacture, then a fault tolerant strategy must be adopted, which allows a chip to work, even with many faulty (either temporarily or permanently) devices. Furthermore, reconfiguration might be very time consuming for protecting against transient errors that may occur in service, and therefore demand temporary shutdown of the system until reconfiguration is performed. It may also be necessary to use NAND multiplexing if reconfiguration methods are impractical or if the probability of transient errors is very high. RMR provides some benefits, but these are unlikely to be useful for chips with 10^{12} devices, once the manufacturing defect rate is greater than about 10^{-8} . The NAND-M technique in principle would allow chips with 10^{12} devices to work, even if the fault rate is as high as 10^{-3} per device. However, this needs even more redundancy than the reconfiguration technique.

The implications of these results are that the future usefulness of various nanoelectronic devices may be seriously limited if they cannot be made in large quantities with a high degree of reliability. The results show that it is theoretically possible to make very large functional circuits, even with one dead device in ten, but only if the dead devices can be located and the circuit reconfigured to avoid them. Even so, this technique would require a redundancy factor of $\sim 10,000$. For example, a chip with 10^{12} non-perfect devices would perform as if it had only 10^8 perfect devices. If it were not possible to locate the dead devices, then one of the other two techniques would have to be used. These would require the manufacturing and lifetime failure rate for $R=1000$ to be between 10^{-7} and 10^{-6} .

BIOLOGICALLY INSPIRED ARCHITECTURE IMPLEMENTATIONS

Biologically inspired computing implies emulation of human and biological reasoning functions. Such architectures possess basic information processing capabilities that are organized and reorganized in goal-directed systems. The living cell is the biological example of a goal-directed organism and has the features of flexibility, adaptability, robustness, autonomy, situation-awareness, and interactivity. The self-organization of biological cells is responsible for their own survival, destruction, replication, and differentiation into multi-cellular forms, all under the direction of goals encoded in their genes. The programming model does not involve millions of lines of code but rather modules of encoded instructions that are activated or deactivated by regulatory modules to act in concert with an overall goal-directed system. Algorithms inspired by computational neurobiology have been the first approach to computing systems that exhibit such

¹⁰⁹ P.G. Depledge, “Fault-tolerant Computer Systems,” *IEEE Proc.* 128, 1981, 257–272.

¹¹⁰ S. Spagocci and T. Fountain, “Fault Rates in Nanochip Devices,” *Electrochem. Soc. Proc.* 99, 1999, 354–368.

¹¹¹ J. Von Neumann, *Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components Automata Studies*, eds. C.E. Shannon and J. McCarthy. Princeton, NJ: Princeton University Press, 1955. 43–98.

¹¹² J.R. Heath, P.J. Kuekes, G.S. Snider and R.S. Williams, “A Defect-tolerant Computer Architecture: Opportunities for Nanotechnology.” *Science* 280, 1998, 1716–1721.

¹¹³ K. Nikolic, A. Sadek, and M. Forshaw, “Fault-tolerant Techniques for Nanocomputers,” *Nanotechnology* 13, 2002, 357–362.

36 Emerging Research Devices

behavior, implemented either as unique processors or on general-purpose architectures. However there is an enormous gap in our understanding of how biological pathways or circuits function. So there is much learning needed before this knowledge can be captured in useable computing systems.

At the nanoscale, devices are more stochastic in operation and quantum effects become the rule rather than the exception. It is unlikely that existing computational models will be an optimal mapping to these new devices and technologies, and this is the motivation for biologically inspired algorithms. Neural circuits use loosely coupled, relatively slow, globally asynchronous, distributed computing with unreliable (and occasionally failing) components. Furthermore, even simple biological systems perform highly sophisticated pattern recognition and control. Biological systems are self-organizing, tolerant of manufacturing defects, and they adapt, rather than being programmed, to their environments. The problems they solve involve the interaction of an organism/system with the real world.¹¹⁴

Biological systems are also inherently low power at these relatively slow speeds. The human brain is known to consume 10–30 W in performing its functions at millisecond timeframes that are compatible with the physiological processes being controlled.

The interconnect capabilities of biologically inspired architectures are the key to its massive parallelism. The connectivity of neurons in humans provides the best-known example of this. One cubic millimeter of cortex contains about 10^5 neurons and 10^9 synapses (10^4 synapses/neuron) and the human nervous system has about 10^{12} neurons and 10^{15} synapses (10^3 synapses/neuron). Thus the fan-out per neuron ranges from 10,000 to 1,000 in humans.¹¹⁵ This amounts to about 1–10 synapses/ μm^3 . Most neurons are not connected to nearest neighbors but rather to different cell classes required to execute the goal-directed function. This enormous interconnectivity requires a much different approach to managing information and algorithmic complexity than we implement in current computing systems. And the large fan-out will require either large-gain devices or circuit approaches based on additional signal processing inputs such as the regulatory enzymes of biological reaction pathways.

The feasibility of using nanoscale electronic devices and interconnects to implement such massively parallel, adaptive, self-organizing computational models is an active research area. In general, such architectures should be of interest for complex digital and intelligent signal processing applications such as advanced human computer interfaces. These interfaces will include elements such as computer recognition of speech, textual, and image content as well as problems such as computer vision and robotic control. These classes of problems require computers to find complex structures and relationships in massive quantities of low-precision, ambiguous, and noisy data.

Implementations of biologically inspired systems can be either entirely analog or digital, or a hybrid of the two. Each has its advantages and disadvantages. Analog has more density than digital, and many of the algorithmic operations, such as leaky integration, that often appear in this class of algorithms, can be implemented very efficiently in analog. Also analog can be much more efficient in terms of power/operation. Digital representation of computations allows more flexibility and allows multiplexing of expensive computer hardware by a number of network nodes. This is particularly attractive when the network is sparsely activated. On the other hand, analog is much harder to design and debug due to the lack of mature design tools. Also analog quantities are much more difficult to store reliably and bit precision may not be acceptable with the small numbers of electrons and low values of voltage and current. Digital implementations use many more transistors and power per operation and must eventually interface with analog signals in the real world.

The communications functions, even in analog systems, are best performed digitally. Most neurons communicate via inter-spike-intervals using the time between pulses to represent a signal versus current or voltage. This type of signaling is very noise tolerant and scales cleanly to single electron systems. Representing addresses in digital forms, such as packets, means that dedicated metal interconnect wires are not required and that the network can grow without adding new wires. Also multiplexing schemes for increasing bandwidth are enabled by digital systems. However single-electron systems do not have the gain required to drive large fan-out circuits typical of biological implementations. Very little work has been performed on nanoscale devices and circuits that would provide such functions.

¹¹⁴ G. Palm et al., "Neural Associative Memories, in *Associative Processing and Processors*," A. Krikelis and C.C. Weems, Editors, IEEE Computer Society, Los Alamitos, CA, 1997, 284–306.

¹¹⁵ Patricia S. Churchland and Terrence J. Sejnowski. *The Computational Brain*. The MIT Press: Boston, MA, 1992. ISBN 0-262-03188-4.

COHERENT QUANTUM COMPUTING

Coherent quantum devices rely on the phase information of quantum wavefunctions to store and manipulate information. The phase information of any quantum state is called a “qubit” and is extremely sensitive to its external environment. It is easily connected or entangled with the quantum states of particles in the local environment. However, no physical system can ever be completely isolated from its environment; the same sensitivity can be used to entangle adjacent qubits in ways that can be controlled by physical gates. The core idea of quantum information processing or quantum computing is that each individual component of an infinite superposition of wavefunctions is manipulated in parallel, thereby achieving massive speed-up relative to conventional computers. The challenge is to manipulate wavefunctions so that they can perform a useful function and then find a way to read out the result of the calculation.

Essentially there have been three approaches for the implementation of quantum computers as follows:

1. Bulk resonance quantum implementations including nuclear magnetic resonance, linear optics, and cavity quantum electrodynamics (CQED)
2. Atomic quantum implementations including trapped ions and optical lattices
3. Solid-state quantum implementations including semiconductors and superconductors

Decoherence is a major issue—where qubits lose their quantum properties (phase information in the wavefunctions) exponentially fast in the presence of a constant amount of noise per qubit. The decoherence per operation ranges from 10^{-3} for electron charge states in semiconductors, to 10^{-9} for photons, 10^{-13} for trapped ions, and 10^{-14} for nuclear spins.⁴⁶ The emphasis of this description is on solid-state implementations with a focus on semiconductors since this is the most attractive for developing the required manufacturing process control and commercial products.

As stated above, the qubit is a fundamental notion in quantum computing, a concept that parallels the “bit” in conventional computation, but carrying with it a much broader set of representations. Rather than a finite dimensional binary representation for information, the qubit is a member of a two-dimensional Hilbert space containing a continuum of elements. Thus quantum computers operate in a much richer space than binary computers. Researchers have defined many sets of elementary quantum gates based on the qubit concept that perform mappings from a set of input quantum registers to a set of output quantum registers. A single gate can entangle the qubits stored in two adjacent quantum registers and combinations of gates can be used to perform more complex computations. It can be shown that, just as in Boolean computation, there exist minimal sets of quantum gates that are complete with respect to the set of computable functions.

Considerable research has been conducted to define the capabilities of quantum computers. Theoretically quantum computers are not inferior to standard computers of similar complexity and speed of operation. More interesting is the fact that for some important classes of problems, the quantum computer is superior to its standard counterpart. In particular, it was shown that the two prime factors of a number can be determined by a quantum computer in time proportional to a polynomial in the number of digits in the number.¹¹⁶ This truly remarkable result showed that for this particular class of problems, the quantum computer is at least exponentially faster than a standard computer. The key to this result is the capability of a quantum computer to efficiently compute the quantum Fourier Transform. This result has immediate application in cryptography since it would allow the quick determination of keys to codes such as RSA. It is estimated that few thousand quantum gates would be sufficient to solve a representative RSA code containing on the order of one hundred digits. There are several other applications that are variants of the factorization problem.¹¹⁷

The development of a practical architecture for reliable quantum computers is just beginning.¹¹⁸ Elementary architecture implementation concepts such as quantum storage, data paths, classical control circuits, parallelism, programming models, and system integration are not yet available. The overhead requirement for quantum error correction is a daunting problem; the error probability for a quantum operation can be as high as 10^{-4} and requires considerable efforts to manage. Improvements in error correction code are in research now but their impact is not yet known. Practical architectures will require error rates between 10^{-6} and 10^{-9} .

¹¹⁶ P.W. Shor, “Algorithms for Quantum Computation: Discrete Logarithms and Factoring,” *Proc. 35th Annual Symposium on Foundations of Computer Science, IEEE Computer Society Press, 1994, 124–134.*

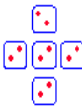
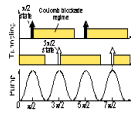
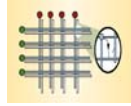
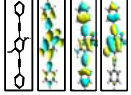

¹¹⁷ C.P. Williams and S. H. Clearwater. *Explorations in Quantum Computing. Springer-Verlag: New York, NY, 1998.*

¹¹⁸ M. Oskin, F.T. Chong, and I.L. Chuang, n.t., *IEEE Computer, January 2002, 79.*

38 Emerging Research Devices

A minimum set of architecture building blocks has been proposed¹¹⁹—a quantum arithmetic logic unit, quantum memory, and a dynamic scheduler. In addition, the architecture implementation uses a novel wiring technique that exploits quantum teleportation. In this wiring, the desired operation is performed simultaneously with the transport.

Table 64 Emerging Research Architecture Implementations

					
Architecture Implementations	Cellular Array Implementations		Defect Tolerant Implementations	Biologically Inspired Implementations	Coherent Quantum Computing
	Quantum Cellular Automata	Cellular Nonlinear Networks			
Application Domain	<ul style="list-style-type: none"> Not demonstrated 	<ul style="list-style-type: none"> Fast image processing Associative memory Complex signal processing 	<ul style="list-style-type: none"> Reliable computing with unreliable devices (such as SETs with background noise) Historical examples include WSI Teramac FPGA implementations 	<ul style="list-style-type: none"> Goal-driven computing using simple and recursive algorithms High computational efficiency through data compression algorithms 	<ul style="list-style-type: none"> Special algorithms such as factoring and deep data searches
Device And Interconnect Implementations	<ul style="list-style-type: none"> Arrays of nanodots or molecular assemblies 	<ul style="list-style-type: none"> Resonant tunneling devices 	<ul style="list-style-type: none"> Molecular switches, Crossed arrays of 1D structures Switchable interconnects 	<ul style="list-style-type: none"> Molecular organic and bio-molecular devices and interconnects 	<ul style="list-style-type: none"> Spin resonance transistors NMR devices Single flux quantum devices Photonics

¹¹⁹ M. Oskin, F.T. Chong, and I.L. Chuang, n.t., *IEEE Computer*, January 2002, 79.

Table 64 Emerging Research Architecture Implementations (continued)

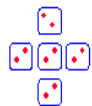
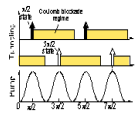
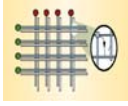
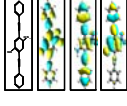

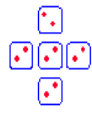
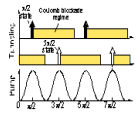
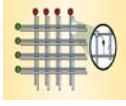
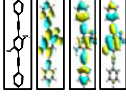

					
Architecture Implementations	Cellular Array Implementations		Defect Tolerant Implementations	Biologically Inspired Implementations-	Coherent Quantum Computing
Desirable Functional Characteristics And Challenges					
Information Throughput	<ul style="list-style-type: none"> Fan out =1 Functional throughput constrained by interdot capacitances 	<ul style="list-style-type: none"> Fan out close to unity 	<ul style="list-style-type: none"> Fan out variable but performance degraded by need for defect management schemes 	<ul style="list-style-type: none"> Massive parallelism Requires some long-range data transfer Fan out very high in brains (max=10⁴ and avg=10³) 	<ul style="list-style-type: none"> Exponential performance scaling Presently limited to 5 qubits but 50–100 qubits needed for large computations
Power	<ul style="list-style-type: none"> Power comparable to scaled CMOS (0.1–0.5 MIPS/mW) Data streaming applications will need 10–100 MOPS/mW 	<ul style="list-style-type: none"> Power comparable to scaled CMOS (0.1–0.5 MIPS/mW) Data streaming applications will need 10–100 MOPS/mW 	<ul style="list-style-type: none"> Not demonstrated 	<ul style="list-style-type: none"> High parallelism results in lower operational speeds Power consumption of human brain 10–30 W at millisecond rates 	<ul style="list-style-type: none"> Not demonstrated for large-scale computations
Interconnects	No local interconnects	Local interconnects with neuron-like waveforms	Interconnects by crossed arrays	Interconnects distributed over a range of distances	Interconnects through wavefunction coupling and entangled states
Error Tolerance	<ul style="list-style-type: none"> Sensitive to background charge Low temperature operation 	<ul style="list-style-type: none"> Not determined 	<ul style="list-style-type: none"> Multiple modular redundancy and multiplexing for transient errors 	<ul style="list-style-type: none"> Highly dynamical neural-like systems Implement adaptive self-organization, fault tolerance 	<ul style="list-style-type: none"> Error correction costs high
Defect Tolerance	<ul style="list-style-type: none"> Not demonstrated 	<ul style="list-style-type: none"> Not determined 	Techniques used include: <ul style="list-style-type: none"> Redundancy NAND multiplexing Reconfiguration 	<ul style="list-style-type: none"> Inherently insensitive to defects through adaptive algorithms 	<ul style="list-style-type: none"> Error correcting algorithms needed
Manufacturability	Precise dimensional control needed	Tight tolerances on tunnel rates of all junctions to minimize jitter	Self assembly possible	Not demonstrated	Demonstrated NMR quantum computing with 6 qubits
Test	Not demonstrated	Demonstrated only for image processing	Self-test or requires extensive pre-computing test	Test functions are included in the adaptive algorithms used	Test not possible directly

Table 64 Emerging Research Architecture Implementations (continued)

					
<i>Architecture Implementations</i>	<i>Cellular Array Implementations</i>		<i>Defect Tolerant Implementations</i>	<i>Biologically Inspired Implementations</i>	<i>Coherent Quantum Computing</i>
<i>Remarks</i>					
<i>Comments</i>	<ul style="list-style-type: none"> ▪ No programming model 	<ul style="list-style-type: none"> ▪ Locally active and locally connected Cell and array design immature (no fan-out) ▪ No programming model 	<ul style="list-style-type: none"> ▪ Supports memory-based computing ▪ Applications in dependable systems 	<ul style="list-style-type: none"> ▪ Goal directed program model ▪ Backed by extensive neural network research ▪ Algorithmic implementations need more research 	<ul style="list-style-type: none"> ▪ Extreme application limitation ▪ No general-purpose architecture or programming model
<i>Maturity</i>	Demonstration	Demonstration	Demonstration	Concept	Concept
<i>Research Activity (2001-2003)</i>	25 research papers	92 research papers	10 research papers	12 research papers	976 research papers

EMERGING TECHNOLOGIES—A FUNCTIONAL COMPARISON

INTRODUCTION

The technological challenges for the information processing industry in the post CMOS-scaling era are quite different because it is not clear what needs to be done. This section relates some of the new information processing technologies to each other and to scaled CMOS using four application-driven parameters to gain an overall perspective on the issues and opportunities.

There is a growing consensus that from about 2019 forward, information-processing technology will consist of a heterogeneous set of novel and widely disparate device technologies integrated on a silicon platform consisting of very fast, very small, and very low-cost CMOS devices. These novel devices will span a very broad range of materials, operational principles, functionalities, logic systems, data representations, and architectures. In general, their characteristics will be complimentary to scaled CMOS, perhaps extending CMOS to new applications. However, none of the new technologies currently being explored is thought to have a real possibility of replacing silicon CMOS.

FUNCTIONAL PARAMETERIZATION AND COMPARISON

Figure 42 shows a parameterization of a selected set of emerging technologies and CMOS in terms of speed, size, cost, and switching energy.¹²⁰ Five of the technologies, including CMOS, are introduced in the CMOS, Logic and Architecture sections and three others (plastic, optical, and NEMS) are described in this section. (Biologically Inspired is further described in this section, and, like CMOS, is plotted for the purpose of reference comparison.) The first three parameters in this figure are used to define a 3D space and the fourth parameter, switching energy, is displayed as color code shown in the legend. All the scales are logarithmic and cover many orders of magnitude as shown in the graph. Each of the technologies displaces a certain volume in this parameter space and is color-coded in a solid color representing the energy required for a single gate operation. Each of the volumes is also projected onto the bounding 2D planes so that quantitative values can be determined. The projections of the volume corresponding to a given technology are shown as rectangles filled with the same color as the corresponding volume.

¹²⁰ Mr. David Jaeger of North Carolina State University is gratefully acknowledged for providing technical support in the preparation of Figure 42.

In the absence of firm measured data, a number of assumptions were made to estimate the parameters for the emerging technologies. The parameters used for each technology are listed in Table 65. If an emerging technology is in the conceptual stage with no measured data, the parametric assumptions are based on the underlying physical principles. If some measured data exists, the assumptions involve an estimate on how far the technology can be scaled. In this case, the scaling arguments are based on physical principles.

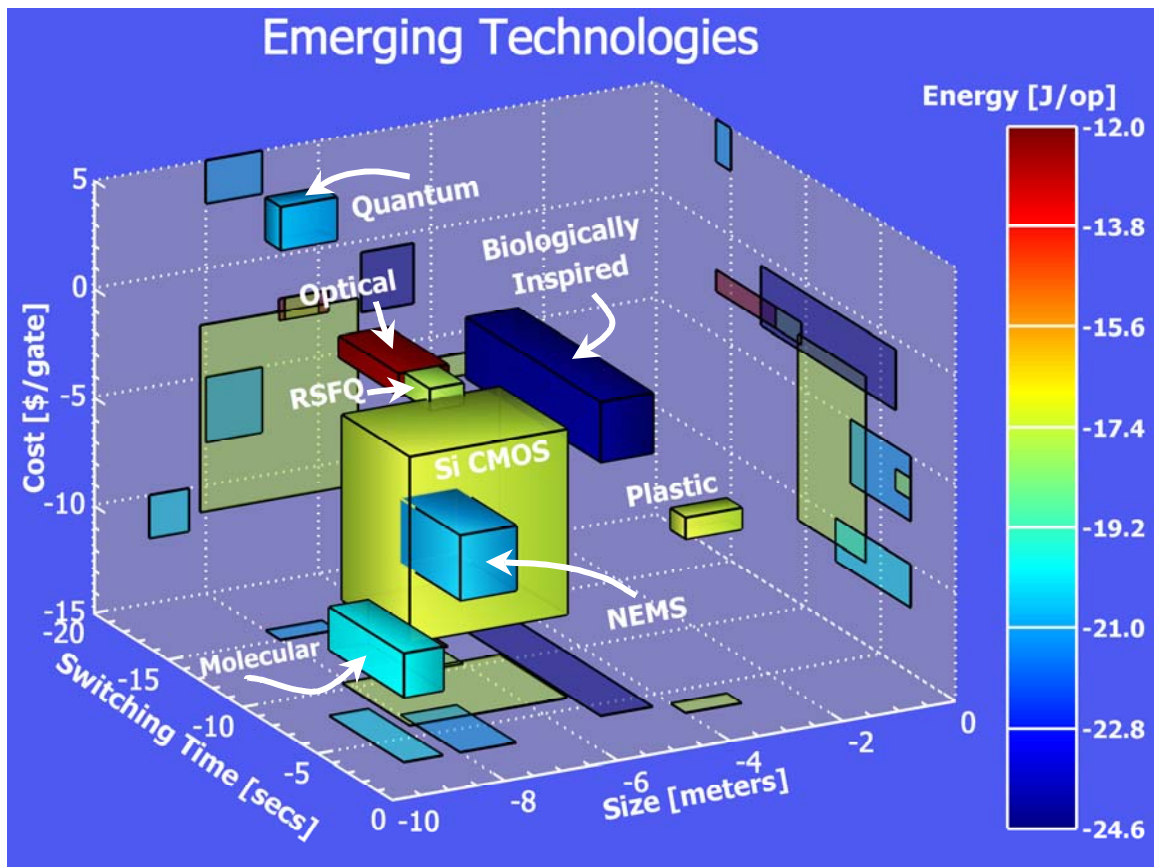


Figure 42 Parameterization of Emerging Technologies and CMOS—Speed, Size, Cost, and Switching Energy

Several of the technologies listed are strongly tied to a single application area or niche where the technology is particularly effective. For example, quantum computing can be used to find prime factors very efficiently by means of Shore's algorithm¹¹⁶ but is much less efficient on other applications. In this case, an "effective" time per operation is defined as the time required by a classical device in a classical architecture using a classical algorithm to do the calculation. Therefore the "effective" operation time of an N-qubit quantum computer factoring a large number is very much faster than the operation time of an N-gate classical computer because of the inherent parallelism associated with quantum computing. Similar arguments can be made for biologically inspired and optical computing.

This figure conveys meaningful information about the relative positions of the emerging technologies in this application space. It shows that few of the new technologies are directly competitive with scaled CMOS and most are highly complimentary. It also shows very clearly the benefit to be derived from heterogeneous integration of the emerging technologies with silicon to expand its overall application space. Figure 42 and Table 65 represent initial estimates for the comparison of these very disparate technologies. In addition to this comparison, the further intent of this figure and table is to stimulate substantive discussion of the basis and means for making this comparison.

Table 65 Estimated Parameters for Emerging Research Devices and Technologies in the year 2016

Technology	T_{\min} sec	T_{\max} sec	CD_{\min} m	CD_{\max} m	Energy J/op	Cost min \$/gate	Cost max \$/gate
Si CMOS	3E-11 ¹²¹	1E-6	8E-9	5E-6	4E-18	1E-11	3E-3
RSFQ	1E-12	5E-11	3E-7	1E-6	2E-18	1E-3	1E-2
Molecular	1E-8	1E-3	1E-9	5E-9	1E-20	1E-12	1E-10
Plastic	1E-4	1E-3	1E-4	1E-3	4E-18	1E-7	1E-6
Optical (digital, all optical)	1E-16	1E-12	2E-7	2E-6	1E-12	1E-3	1E-2
NEMS (conservative)	1E-7	1E-3	1E-8	1E-7	1E-21	1E-8 ¹²²	1E-5
Biologically Inspired	1E-13	1E-4	6E-6	5E-5	3E-25	5E-4	3E-1
Quantum	1E-16	1E-15	1E-8	1E-7	1E-21	1E3	1E5

In this table T stands for system cycle time (switching time), CD stands for critical dimension (e.g., physical gate length), Energy is the intrinsic operational energy of one device, and Cost is defined as \$ per gate.

DEFINITION AND DISCUSSION OF TABLE ENTRIES

Plastic Transistors—Plastic transistors are defined to be thin film transistor (TFT) devices fabricated on plastic substrates. The active layer of the TFT can be amorphous or poly-Si as well as organic semiconductors. Often, the TFTs are combined with organic light emitting diodes (OLEDs) to form intelligent, flexible display devices than can be bent, folded, worn or conformally mapped on to arbitrarily shaped surfaces. All-plastic chips based entirely on organic materials have already been demonstrated whose mechanical flexibility offers totally new perspectives to, for example, the rapidly growing market of identification and product tagging as well as for pixel drivers for flexible displays.¹²³ Typical devices have a supply voltage of 10 V and critical dimensions of 100 μm with reasonable electron mobilities and I-V characteristics. Pentacene-based plastic transistors with $I_{\text{on}}/I_{\text{off}}$ current ratio $>10^5$ at operating voltage ranges as low as 5 Volts have been reported.¹²⁴ Analog and digital circuits using organic (pentacene) transistors on polyester substrates have been fabricated and characterized. The highest operation frequency reported to date for organic circuits on plastic substrates is 1.7 kHz.¹²⁵ Plastic transistors have the potential to provide very low-cost, rugged large area electronics that have many potential applications.^{126, 127} A process technology consisting just of printing operations on paper-based substrates would have an intrinsic cost structure similar to color inkjet printing today. It could support disposable devices such as periodicals and dynamic bar codes.

¹²¹ T_{\min} for silicon CMOS is based on the local clock rate for the 22 nm node (physical gate length < 9 nm), and not upon CV/I intrinsic switching time.

¹²² Estimated on the principle of reasonable cost and assumed two-dimensional architecture of NEMS computer.

¹²³ F. Würthner, "Plastic Transistors Reach Maturity for Mass Applications in Microelectronics," *Angew. Chem. Int. Ed.* 40, 2001, 1037–1039.

¹²⁴ C.D. Dimitrakopoulos, S. Purushothaman, J. Kymissis, A. Callegari, J. M. Shaw, "Low-voltage Organic Transistors on Plastic Comprising High-dielectric Constant Gate Insulators," *Science* 283, 1999, 822–824.

¹²⁵ M.G. Kane, J. Campi, M.S. Hammond, F.P. Cuomo, B. Greening, C.D. Sheraw, J.A. Nichols, D.J. Gundlach, J.R. Huang, C.C. Kuo, L. Jia, H. Klauk, T.N. Jackson, "Analog and Digital Circuits using Organic Thin-Film Transistors on Polyester Substrates," *IEEE Electron. Dev. Lett.* 21, 2000, 534–536.

¹²⁶ J.M. Xu, "Plastic Electronics and Future Trends in Microelectronics," *Synthetic Metals* 115, 2000, 1–3.

¹²⁷ S. Forrest, P. Burrows, and M. Thompson, "The Dawn of Organic Electronics," *IEEE Spectrum*, August 2000, 29–34.

*Optical*¹²⁸—Optical computing is based on using light transmission and interaction with solids for information processing. The potential advantages of digital optical computers relate to the following properties of light as a carrier of information:

- Optical beams do not interact with each other
- Optical information processing functions can be performed in parallel (e.g., performing a Fourier transform)
- Ultimate high speed of signal propagation (speed of light)

It should be noted that what is called the all-optical computer still contains electronic components, such as lasers and nonlinear elements in which a material's optical properties are affected by charge carriers or atoms interacting with light. Some disadvantages of digital optical computing include the following:

- The relatively large size of components (e.g., optical switch) arising from diffraction limitation
- Potential of high-speed computation can be realized only at the expense of dissipated power. For example, in an optically controlled phase change material (switch or memory), faster rearrangement of atoms in a cell requires a larger supply of energy. In a practical device “computing at the speed of light” is unlikely since it would require a huge operational energy.

Near-term opportunities in optoelectronics are in integration of photonic components with sub-100 nm CMOS. Another opportunity arises from using optically controlled phase-change materials, such as PCM described in the Memory section. Another direction is perfection of existing analog optical computers, which perform Fourier processing much faster than electronics. Analog optical computers are fast and operate with continuous data, while their accuracy is not comparable to that of digital computers.

Nano-electro-mechanical systems (NEMS)—In the concept of the nanomechanical computer, mechanical digital signals are represented by displacements of solid rods, and the speed of signal propagation is limited to the speed of the sound (for example, 1.7×10^4 m/s in diamond). Optimistic estimates predict NEMS logic gates that switch in 0.1 ns and dissipate less than 10^{-21} J and computers that perform 10^{16} instructions per Watt (compared to 5×10^{12} instruction per Watt in human brain operation). This estimated switching energy is below the thermodynamic limit of $kT \ln 2$ for irreversible computation. It is believed¹²⁹ that this low dissipation is possible because NEMS computation is logically reversible. More conservative estimates of characteristics of the NEMS computers can be made based on recent demonstration of a VLSI-NEMS chip for parallel data storage.¹³⁰ Reported storage densities are 500 Gbit/in². The highest data rates achieved so far are 6 Mbit/sec. A summary of the conservative estimates of parameters of the NEMS computers is given in Table 65.

Biologically Inspired—The human brain is defined to be the archetypal *biologically inspired* or *neuromorphic* information processing device and is included here to provide a basis of comparison with silicon-based information processing systems. The scale length of individual neurons is estimated from the volume of the brain and the estimated number of neurons. It is possible to derive an “effective operation time” of biologically inspired computing as explained in the overview of this section. In that case, the reference operation is vision processing where there is a great deal of information relating to technological systems. The effective times defined in this way are very much faster than the synaptic speed and reflects that the interconnect density of the human brain is very much greater than any silicon-based system. The speed quoted in Table 65 for T_{\min} is based on the estimated information-processing rate of 1×10^{13} bits per second¹³¹ related to vision processing. Similarly, the speed quoted in Table 65 for T_{\max} is the experimentally observed time scale for opening and closing of synapses. Each neuron will connect to between 100 and 10,000 synapses, one of the primary ways in which the architecture of the human brain differs from silicon-based systems.

¹²⁸ H.J. Caulfield, “Perspectives in Optical Computing,” *Computer*, February 1998, 22–25.

¹²⁹ K. Eric Drexler. *Nanosystems: Molecular Machinery, Manufacturing and Computation*. John Wiley & Sons, Inc.: New York, NY, 1992.

¹³⁰ M. Despont, J. Brugger, U. Drechsler, U. Düring, W. Haberle, M. Lutwyche, H. Rothuizen, R. Stutz, R. Widmer, G. Binnig, H. Rohrer, P. Vettiger, “VLSI-NEMS Chip for Parallel AFM Data Storage,” *Sensors and Actuators* 80, 2000, 100–107.

¹³¹ R.U. Ayres. *Information, Entropy, and Progress*. AIP Press: New York, NY, 1994.

44 Emerging Research Devices

The fundamental parameters of the human brain¹³² are estimated to be:

- Number of neurons—2E10
- A single neuron can make 100 to 10,000 synaptic connections
- Mass—1.3 kg¹³³
- Volume—600 cm³
- Power consumption—15–30 Watts
- Information stored—1E12 (short term) bits
- Information processed—1E16 bits/second

The set of secondary parameters shown in Table 65 is based on the fundamental parameters above.

EMERGING TECHNOLOGIES—A CRITICAL REVIEW

INTRODUCTION

While the role of nanoscale devices in meeting future computing and communications applications is not clear at this point, undoubtedly there will be many needs that could benefit from the terascale level of integration that such devices offer. As discussed in the previous section, these devices will encompass a broad range of fabrication methodologies and functional modalities. They may extend scaled CMOS to new applications in a highly complementary fashion. Conversely, there are significant limitations that arise with nanoscale devices that will impact their usefulness. In particular, as mentioned above, their near-term applications will require nanoscale devices to be functionally and technologically compatible with silicon CMOS. In the longer term, charge-based nanoscale devices may be supplemented with one or more new information processing technologies using a quite new logic “state variable” or means of representing the bit. The purpose of this section, therefore, is to introduce a set of technology evaluation criteria (see notes for Table 66) and, based on these criteria, to offer a critical assessment of those technology entries for memory and logic being considered for post CMOS-scaling information processing. Additionally, charge-based approaches will be discussed in this section separately from those approaches proposing use of a new means for data representation or “state variable.” This separate discussion addresses an important question related to new charge-based information processing approaches concerning the fundamental limits of an elemental switch (size, energy, speed, etc.).

TECHNOLOGIES BEYOND CMOS

OVERALL TECHNOLOGY REQUIREMENTS

Gain—The gain of nanodevices is an important limitation for current combinatorial logic where gate fan-outs require significant drive current and low voltages make gates more noise sensitive. New logic and low-fan-out memory circuit approaches will be needed to use most of these devices for computing applications. Signal regeneration for large circuits may need to be accomplished by integration with CMOS. In the near-term integratability of nanodevices to CMOS silicon is a key requirement due to both the need for signal restoration for many logic implementations and also the established technology and market base. This integration will be necessary at all levels from design tools and circuits to process technology.

Power limitations—Clock speed versus density trade-offs for electron transport devices will dictate that for future technology generations, clock speed will need to be decreased for very high densities or conversely, density will need to be decreased for very high clock speeds. In other words, the power-delay product (minimum power dissipated \times (switching time)²) cannot be less than Planck’s constant, h , in the quantum limit. Nanoscale electron transport devices mostly fit into the former category and will best suit implementations that rely on the efficient use of parallel processing more than on fast switching.

Device transfer function—Nanoscale devices may perform circuit functions directly due to their nonlinear outputs and therefore save both real estate and power. In addition, nanodevices that implement both logic and storage in the same device would revolutionize circuit and nanoarchitecture implementations.

¹³² *Nanoelectronics and Information Technology*, ed. Rainer Waser. Wiley-VCH, 2003. 350.

¹³³ L.C. Aiello and P. Wheeler, “The Expensive-tissue Hypothesis: The Brain and the Digestive System in Human and Primate Evolution,” *Curr. Anthropology* 36, 1995, 199–221.

Output impedance and contact resistances—The total output impedance for electron transport devices can be more than 100 k Ω so that, for comparable interconnect capacitances, the lowest impedance device will be favored. In fact, device output impedance may be even more important than interconnect resistance for nanoscale devices. This is reflected in the low gain of these devices. Many nanoscale devices have output impedance of hundreds of M Ω s or more. Contact resistances for metal nanodevice contacts must be much less than the channel or the device operating region resistance, comparable to the interconnect resistances, and they must be repeatable and reliable.

Error rate—The error rate of all nanoscale devices and circuits is a major concern. These errors arise from the highly precise dimensional control needed to fabricate the devices and also from interference from the local environment, such as spurious background charges in SETs. It has been estimated that redundancies of 10^3 to 10^4 will be needed for manufacturing and lifetime device failure rates of 10^{-6} to 10^{-7} . Thus for nanodevice levels of 10^{12} , only 10^8 to 10^9 devices will be useable for computation. Large-scale error detection and correction will need to be a central theme of any architecture and implementations that use nanoscale devices. (Refer to Defect Tolerant Implementation in the Architectural section.)

Operation temperature—Nanodevices must be able to operate at or close to room temperature for practical applications.

Interconnect limitations—Nanodevices based on electron transport must be interconnectable without a major loss in density, performance, or power. Interconnects must demonstrate transmission resistances of several tens of k Ω . The interconnect pitch transformation from nanoscale dimensions to the order of millimeters used in most applications will require sophisticated multiplexing schemes to enable bi-directional signal flows.

CHARGE-BASED NANOSCALE DEVICES

An important issue regarding charge-based nanoelectronic switch elements is related to the fundamental limits to the scaling of these new devices, and how they compare with CMOS technology at its projected end of scaling. The 2003 ITRS projects the scaling of CMOS to the 22 nm node or even below. This node represents a physical gate length for an MPU/ASIC device of 9 nm with an average power dissipation of 93 W/cm². A recent analysis¹³⁴ concludes that the fundamental limit of scaling a charge-based switch is only a factor of 5–10 \times smaller than the physical gate length of a CMOS MOSFET in 2018. Furthermore the density of these switches is limited by maximum allowable power dissipation of approximately 100 W/cm², and not by their size. The conclusion of this work is that MOSFET technology scaled to its practical limit in terms of size and power density will closely reach the theoretical limits of scaling for charge-based devices. Consequently, application of *emerging charge-based* logic technologies, such as 1D structures (nanowires and nanotubes) and molecular structures may be best suited for use as a replacement of the silicon channel in an otherwise silicon-based MOSFET technology infrastructure. In other words, use of 1D or molecular structures for *charge-based switches* to develop a completely new information processing technology, including binary switches, memory elements, interconnects (local and global) may not be justified to obtain a relatively modest maximum of 5-10 \times scaling in size or speed. This conclusion is particularly true since the device density is limited by power dissipation and not by the size of the binary switch. The corollary of this observation is that the search for alternative logic devices should embrace the concept of using state variables other than electric charge.

ALTERNATE LOGIC-STATE-VARIABLE NANOSCALE DEVICES

In this context, the term “state variable” refers to the notion of the finite state machine introduced by Turing in 1930s. The idea is that there are numerous ways to manipulate and store computational information or logic state. The earliest example of a finite state storage device was the abacus, which represents numerical data by the position of beads on a string. In this example, the state variable is simply a physical position, and the operator accomplishes readout by looking at the abacus. The operator's fingers physically move the beads to perform the data manipulations. Early core memories used the orientation of magnetic dipoles to store state. Similarly, paper tapes and punch cards used the presence or absence of holes to store state. Two examples of more recent research activities in alternative state variables for logic are described below.

¹³⁴ V.V. Zhirnov, R.K. Cavin, J.A. Hutchby, G.I. Bourianoff, “Limits to Binary Logic Switch Scaling—A Gedanken Model,” *Proc. IEEE*, November 2003, 1934–1939.

Spin-based devices—New nanoscale devices, such as quantum dots and single-atomic or nuclear spins, based on quantum electron behavior rather than electron transport, offer significant relief from the thermal dissipation problems of electron transport devices. However problems of manufacturability and low-temperature operation are major obstacles to early implementation for quantum dot structures. For molecular structures, operation speed and defect tolerances remain to be explored.

Coherent quantum computing—Coherent quantum effects using devices based on photons, electron or nuclear spins, and superconducting devices will require new computing structures. Such devices function by superposed wave functions that are entangled as qubits and that easily decohere when interacting with an external environment, such as a measurement device. Although enormously capable for a few selected algorithms, such as encryption or deep database searching, quantum computing is not seen yet as being of more general interest. Also the nanofabrication requirements for arrays of coherent quantum devices needed to maintain the needed coherent states are far beyond any extrapolated process control capabilities. So, for now, there is no known path to produce quantum-computing systems with more than a few qubits. Substantial new research will be required to invent, explore, and develop a manufacturable semiconductor-based approach to realize a commercially viable quantum computing systems technology.

POTENTIAL PERFORMANCE AND RISK ASSESSMENT FOR MEMORY AND LOGIC DEVICES

The purpose of this section is to assess the combined potential performance and projected risk associated with each new memory and logic nanoscale device technology discussed for post-CMOS scaling applications in this section. This potential/risk assessment can help inform industrial evaluation of each nanoscale device technology and the industry's investment decisions among the many competing approaches. The Relevance Criteria and Technology Performance Risk Evaluation are given and defined below.

RELEVANCE CRITERIA

Introduction to Criteria—Post CMOS-scaling nanoscale devices span multiple applications, state variables, and technologies and are extremely diverse in nature. A set of nanoelectronics relevance conditions has been defined to parameterize the extent to which a given technology is applicable to information processing applications, particularly those in the near term.

Each post CMOS-scaling nanoscale memory and logic technology is evaluated against each Relevance Criteria according to two factors. The first factor relates to the *projected potential performance* of each nanoscale device technology, assuming its successful development to maturity, for each Relevance Criteria, *compared to that for silicon CMOS at the 22 nm node*. Performance potential is assigned a value from 1–3, with “3” substantially exceeding CMOS at the 22 nm node, and “1” substantially inferior to CMOS (see Factor 1 below). The second factor relates to the risk projected for each technology of achieving its potential performance related to each Relevance Criteria. Again, a numerical scheme is used to assess risk, with risk being assigned a value from 1 to 3. A value of “3” is used for lowest risk and a value of “1” is used for highest risk (see Factor 2 below). The total evaluation is the convolution of these two factors (see Overall Performance and Risk Assessment below). The Relevance Criteria are defined in the notes of Tables 66 and 67.

Factor 1 Individual Performance Potential for each Technology Evaluation Criterion

3	Substantially exceeds CMOS * <i>or</i> is compatible with CMOS architecture ** <i>or</i> is monolithically integrable with CMOS wafer technology *** <i>or</i> is compatible with CMOS operating temperature
2	Comparable to CMOS * <i>or</i> can be integrated with CMOS architecture with some difficulty ** <i>or</i> is functionally integrable (easily) with CMOS wafer technology *** <i>or</i> requires a modest cooling technology, $T \geq 77K$
1	Substantially (2x) inferior to CMOS * <i>or</i> can not be integrated with CMOS architecture ** <i>or</i> is not integrable with CMOS wafer technology *** <i>or</i> requires very aggressive cooling technology, $T < 4K$

Factor 2 Individual Risk Assessment for each Technology Evaluation Criterion

3	Solutions to accomplish most of the Technology Evaluation Criteria for the Technology Entry are known resulting in lowest risk
2	Concepts to accomplish most of the Technology Evaluation Criteria have been proposed for the Technology Entry and are judged to be of moderate risk
1	No solutions or concepts have been proposed accomplish most of the Technology Evaluation Criteria for the Technology Entry and are judged to be of highest risk

Overall Performance and Risk Assessment (OPRA) = Sum [(Performance Potential) x (Risk Assessment)]
(Summed over the eight Evaluation Criteria for each Technology Entry)

Maximum Overall Performance and Risk Assessment (OPRA) = 72

Minimum Overall Performance and Risk Assessment (OPRA) = 8

Overall Performance and Risk Assessment for Technology Entries

Potential for the Technology Entry is projected to be significantly better than silicon CMOS (compared using the Technology Evaluation Criteria) and solutions to accomplish the most of the Technology Evaluation Criteria are known resulting in lowest risk (OPRA ≥ 50)	Potential/Risk
Potential for the Technology Entry is projected to be comparable to or slightly less than silicon CMOS (compared using the Technology Evaluation Criteria) and concepts to accomplish most of the Technology Evaluation Criteria have been proposed and are judged to be of moderate risk (OPRA = 40 – 49)	Potential/Risk
Potential for the Technology Entry is projected to be comparable to or less than silicon CMOS (compared using the Technology Evaluation Criteria) and concepts to accomplish a few of the Technology Evaluation Criteria have been proposed and are judged to be of higher risk (OPRA = 30 – 39)	Potential/Risk
Potential for the Technology Entry is projected to be significantly less than silicon CMOS (compared using the Technology Evaluation Criteria) and no solutions or concepts have been proposed accomplish most of the Technology Evaluation Criteria and are judged to be of highest risk (OPRA < 30)	Potential/Risk

Table 66 Technology Performance and Risk Evaluation for Emerging Research Memory Device Technologies (Potential/Risk)

Memory Device Technologies (Potential/Risk)	Performance [A]	Architecture compatible [B]*	Stability and reliability [C]	CMOS compatible [D]**	Operate temp [E]***	Energy efficiency [F]	Sensitivity $\Delta(\text{parameter})$ [G]	Scalability [H]
Floating Body DRAM	2.3/2.3	3.0/3.0	2.0/2.7	3.0/3.0	3.0/3.0	2.0/3.0	2.3/2.9	2.8/2.7
Phase Change Memory	2.6/2.9	2.2/3.0	2.3/2.2	2.2/3.0	3.0/3.0	1.8/2.7	2.1/2.1	2.7/2.2
Nano-floating Gate Memory	3.0/2.2	2.9/3.0	2.0/2.7	3.0/3.0	3.0/3.0	2.1/2.8	1.6/2.0	2.4/2.0
Insulator Resistance Change Memory	2.4/2.1	2.7/2.7	2.2/2.4	2.1/2.8	3.0/2.9	2.8/2.0	2.1/2.0	2.7/2.4
Molecular Memory	1.6/1.2	1.8/2.0	1.8/1.4	1.9/2.1	2.8/2.3	2.3/1.9	2.1/1.7	2.6/2.2
Single/Few Electron Memory	1.1/1.3	1.9/1.3	1.1/1.0	2.4/1.9	1.3/1.3	2.4/1.2	1.3/1.0	2.6/1.4

Table 67 Technology Performance and Risk Evaluation for Emerging Research Logic Device Technologies (Potential/Risk)

Logic Device Technologies (Potential/Risk)	Performance [A]	Architecture compatible [B]*	Stability and reliability [C]	CMOS compatible [D]**	Operate temp [E]***	Energy efficiency [F]	Sensitivity $\Delta(\text{parameter})$ [G]	Scalability [H]
1D Structures	2.3/2.2	2.2/2.9	1.9/1.2	2.3/2.4	2.9/2.9	2.6/2.1	2.6/2.1	2.3/1.6
RSFQ Devices	2.7/3.0	1.9/2.7	2.2/2.8	1.6/2.2	1.1/2.7	1.6/2.3	1.9/2.8	1.0/2.1
Resonant Tunneling Devices	2.6/2.0	2.1/2.2	2.0/1.4	2.3/2.2	2.2/2.4	2.4/2.1	1.4/1.4	2.0/2.0
Molecular Devices	1.7/1.3	1.8/1.4	1.6/1.4	2.0/1.6	2.3/2.4	2.6/1.3	2.0/1.4	2.6/1.3
Spin Transistor	2.2/1.7	1.7/1.6	1.7/1.7	1.9/1.4	1.6/2.0	2.3/2.1	1.4/1.7	2.0/1.4
SETs	1.1/1.2	1.7/1.2	1.3/1.1	2.1/1.4	1.2/1.8	2.6/2.0	1.0/1.0	2.1/1.7
QCA Devices	1.4/1.3	1.2/1.1	1.7/1.8	1.4/1.6	1.2/1.4	2.4/1.7	1.6/1.1	2.0/1.4

Relevance Criteria Notes for Tables 66 and 67:

[A] Performance—Future performance metrics will be very similar to current performance metrics. They are cost, size, speed and energy dissipation.

[B] Architectural compatibility—This criterion is motivated by the same set of concerns that motivate the CMOS compatibility, namely the ability to utilize the existing CMOS infrastructure that currently exists. The architectural compatibility is defined in terms of the logic system and data representation used by the alternative technology. CMOS utilizes Boolean logic and a binary data representation and ideally, the alternative technology would need to do so as well.

[C] Stability and reliability—As devices approach the atomic scale, structural compositional stability to thermal fluctuations becomes a significant concern. Any realistic alternative device must show structural stability at room temperature for at least 7 years.

[D] CMOS compatibility—The semiconductor industry has been based for the last 40 years on incremental scaling of device dimensions to achieve performance gains. The principle economic benefit of such an approach is it allows the industry to fully apply previous technology investments to future products. Any alternative technology will need to utilize the tremendous investment in infrastructure to the highest degree possible.

[E] Room temperature operation—Room temperature operation is desirable because advanced cooling systems can add substantially to the cost.

[F] Energy efficiency—Energy efficiency appears likely to be the limiting factor of any post CMOS device using electric charge or electric current as a state variable. It also appears likely that it will be dominant criterion in determining the ultimate applicability of alternate state variable devices.

[G] Sensitivity to parametric variation—As devices approach the atomic scale, they become very sensitive to manufacturing and environmental variations. Thus parametric sensitivity is an important criterion for evaluation of alternative technologies. The goal should be a device that is affected but not dominated by parametric variations.

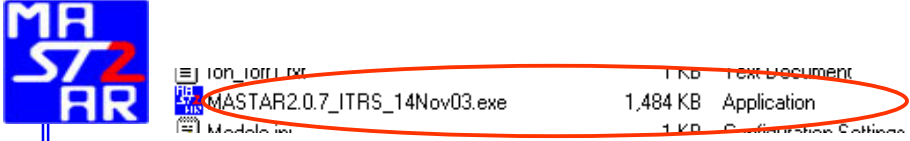
[H] Scalability—In order to derive the economic benefit of incrementalism, any alternative technology should be scalable through multiple generations. It will be desirable to make incremental modifications to the alternative technology and achieve integer multiples of performance. In other words, it should be possible to articulate a Moore's law for the proposed technology.

APPENDIX MASTAR

For the MASTAR general instructions, [click here](#).

The MASTAR application must be downloaded to your hard drive and installed there in order to run it.

1. Put your mouse pointer on the icon for MASTAR and click to launch a page in Internet Explorer or in Netscape.
Your browser may ask if it is okay to open or launch the page.
2. The file to download is called MASTAR.ZIP.
3. Save this file to your hard drive in a new folder and unzip it there, using the option to "unzip to folder."
4. Double-click on the file with the MASTAR icon to run the application.



The screenshot shows a file list with three entries. The second entry, 'MASTAR2.0.7_ITRS_14Nov03.exe', is circled in red. To the left of the file list is the MASTAR logo, which consists of the letters 'MA', 'STAR', and 'AR' stacked vertically in a blue square with white text.

ion_tor1.txt	1 KB	Text Document
MASTAR2.0.7_ITRS_14Nov03.exe	1,484 KB	Application
Modelo.in	1 KB	Configuration Settings