♦ The Future of Solid-State Electronics

William F. Brinkman and Mark R. Pinto

For more than thirty years, the capability of the integrated circuit (IC)—whether it is memory size, processor speed, or cost per transistor—has increased at an exponential rate. This capability increase was achieved through a steady stream of technical innovations in physical sciences, design techniques, and manufacturing methods. However, serious challenges to continued scaling, such as the limits of optical lithography and the complexity of wiring, loom on the horizon. This paper explores the direction in which IC technology is headed, highlights potential roadblocks and possible solutions, and discusses some of the physical considerations that could determine the ultimate limits of integration.

Introduction

Continuous advances in semiconductor technology have precipitated the exponential improvements in the capability of the integrated circuit (IC). While evidence suggests that scaling of complementary metal-oxide semiconductor (CMOS) devices can continue for another ten to fifteen years—down to dimensions on the order of 0.05 μm or 50 nm—some interesting hurdles must be surmounted for this to occur. Eventually, hard physical laws will be encountered that will limit progress, but it is also possible that economic or design complexity limits may surface even sooner. The purpose of this paper is to discuss how the technologies behind the IC will evolve in the future and what the most likely limiting factors will be. While the technology behind making ever-smaller transistors is the principal focus, other technical challenges also are highlighted—for example, manufacturing, design, and multilevel interconnection.

Even before examining any specific IC structures, a number of fundamental physical requirements and limits can be clearly identified.[1] Information must be processed quickly, but signal propagation is limited by transmission line characteristics—for example, it will take ≈ 0.3 ns, at best, for an electrical signal to propagate across a 10-cm^2 chip. Power dissipation must be limited to avoid exceeding the maximum operating temperatures associated with constitutive materials (also related to the ability of the package and environment to remove heat), but thermodynamics and quantum mechanics specify minimum bounds on the energy dissipated by a logic operation—for example, of order kT for irreversible processes. Devices must be organized in a manner that prevents additive propagation of noise, thereby requiring minimal signal levels of several kT/q and finite power gain at each stage. In addition, devices and circuits must operate reliability over some reasonable length of time—typically ten years—in a variety of environmental conditions.

Manufacturing issues begin with limits on the ability to physically define the smallest features. However small the ultimate devices are, they will be worth little unless they can be produced by the billions, thereby placing severe limitations on the variability of each device, as well as on process control. Most importantly, the technology must be economic—for example, a 1-Gb memory must be less expensive than four 256-Mb memories in a module. If process costs escalate too quickly in going to a new generation—that is, via complex steps or much reduced throughput—it is unlikely that progress will continue. Finally, even if physics and manufacturing permit continued improvements, progress in the capability of the IC may also be limited by the exponentially increasing complexity of

design. All these restrictions must be satisfied by any alternative to CMOS technology. Clearly, no such technology exists today. However, it seems foolhardy to believe that human invention will cease when we enter the twenty-first century.

In this paper, we first review the characteristics of the Moore Plot, which has provided a representation of the progress in ICs for more than thirty years. Next, we discuss some of the key challenges that will be encountered in delivering on this plot for the next ten to fifteen years, with lithography clearly being the principal process issue the industry faces. We then look in depth at CMOS device technology and consider its limitations—exaggerated by process variations and yield requirements on billion-transistor circuits—and potential alternatives to CMOS. Issues related to interconnection as one scales to small devices are highlighted. We discuss how the limitations arising from both the device and interconnect technologies may be overcome by innovations in design techniques and CAD. We conclude with some comments about possible directions, anticipating the more detailed presentations on this subject in other papers in this issue.

Living Out the Moore Curve

One of the remarkable aspects of the semiconductor industry is that it has been able to remain on the so-called Moore Plot, first proposed by Gordon Moore in 1965, whereby the memory capacity of a single chip has increased exponentially, currently by a factor of 4, every three years.[2] This steady progress is not due solely to the fact that the minimum feature size has decreased. Between each generation, the minimum dimension has actually changed only by a factor of $\sqrt{2}$, resulting in a chip density increase of only 2. The additional factor of 2 comes equally from increasing the size of chips and innovations in all aspects of the process and design.

There are many examples representing the kinds of innovations that have allowed the aforementioned progress. Perhaps the most obvious is the evolution of the capacitor of the dynamic random access memory (DRAM) cell, which has gone from an essentially horizontal structure to a vertical cylindrical structure that is mounted atop the access transistor. Another example is the industry-wide move to multilevel metallization. It is clear that this innovation has allowed a considerable increase in density. Finally, there are many innovations in design at both the gate and cell levels, as well as at the algorithm level, that have increased the functionality of the circuits produced. **Figure 1** plots the contribution of each of these components to overall progress. Any one component would not have gotten us to where we are today. Clemens provides a detailed discussion on the nature of today's technology in the paper following this one.[3]

Panel 1. Abbreviations, Acronyms, and Terms

AFM—atomic force microscopy
ASIC—application-specific integrated circuit
BiCMOS—bipolar CMOS
CHINT—charge injection transistor
CMOS—complementary metal-oxide semiconductor
DIBL—drain-induced barrier lowering
DRAM—dynamic random access memory
DUV—deep ultraviolet
e-beam—electron beam
EEPROM—electrically erasable programmable read-only memory
EM—electromigration
EPROM—erasable programmable read-only memory
EUV—extreme ultraviolet
IC—integrated circuit
MOS—metal-oxide semiconductor
MOSFET—metal-oxide semiconductor field-effect transistor
NMOSFET—n-type MOSFET
PADRE—PISCES And Device REplacement, a Bell Labs semiconductor simulation program
PMOSFET—p-type MOSFET
PMOS—p-type metal-oxide semiconductor
RC—resistance-capacitance product that defines a signal delay
RF—radio frequency
RTT—resident tunneling transistors
SET—single-electron transistor
SIA—Semiconductor Industry Association
SOI—silicon-on-insulator
STM—scanning tunneling microscopy
VLSI—very large scale integration
W—tungsten

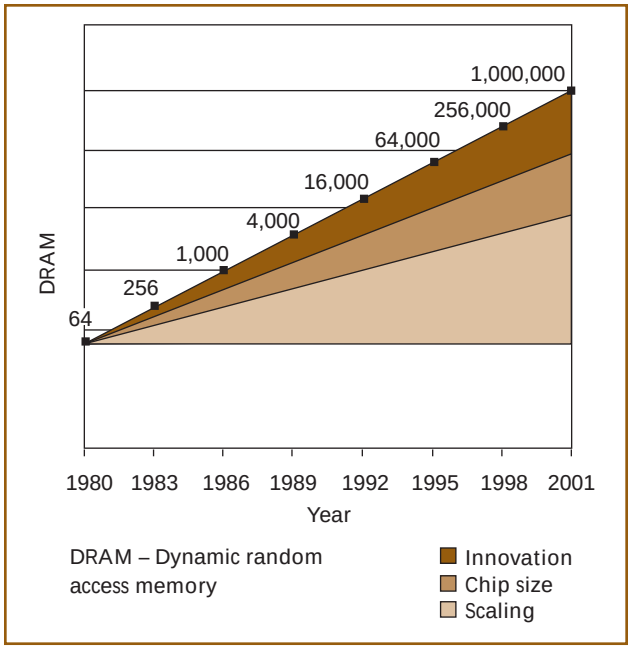

Figure 1.
Moore Plot showing the contribution of design innovation to the overall progress of IC fabrication.

None of these innovations would have mattered if increased integration had come at a higher price. In fact, the process cost of a square centimeter of a silicon IC is approximately the same as it was in the 1970s. This can be attributed to many factors. Clearly, however, the increasing size of wafers and the continuous drive toward clean high-yield manufacturing have had major impacts. Increasing the wafer's size has a huge effect on cost because the time taken per process step is relatively independent of size. Thus, switching from 6" to 8" wafers decreases costs almost by a factor of 2, counterbalancing the increased process and equipment costs due to decreasing line rules.

To what extent the industry can continue moving along this path is an open question. The cost of developing a complete set of equipment for working with larger wafers is becoming so great that no one company can afford to lead the transition to 12" wafers. Because of this, the industry has set up consortia in the United States and Japan to address this issue. The increased output of an optimally sized facility and the associated expense has also stimulated the formation of numerous joint manufacturing ventures, sometimes among firms competing in the same applications markets.

IC yield is affected by two primary factors: aerial defects and statistical component variations. To keep the cost of silicon constant, the industry has had to progress substantially in both areas, though inevitably through more expensive equipment. For instance, with respect to defects, both the density, D_0, and physical size have had to decrease. Currently, it is necessary to control defects down to 0.125 μm and $D_0 < 0.1$ cm^{-2}. Further, to realize a return on the huge investments in new fabrication facilities as quickly as possible, the rate of improvement of D_0 with volume ramp-up has had to improve dramatically. As circuits are made with a larger number of devices, the effect of component variations—in active and passive devices, as well as in interconnections and packaging—may become the most dominant concern.

Mathematically, overall yield, Y_{IC}, can be expressed as a function of the probabilities due to defects Y_D and process variations Y_{VAR} as

$$Y_{IC} = Y_D (D_0, A) \times [\, Y_{VAR} (\sigma, \Delta) \,]^n.$$

The yield due to defects alone depends only on the density, D_0, and on the area of the chip A and not on the number of components, n. Y_{VAR} is the probability of individual components being in specification, and it depends on the manufacturing variation, σ, in electrical properties and the respective margin, Δ, that can be tolerated by the circuit design. Each component must be in specification for the IC to work—hence, the power dependence on n. With $n > 10^9$, the probability Y_{VAR} need not be small to drive Y_{IC} to zero. For example, assuming a normal distribution of some electrical property, the manufacturing variation, σ, would have to be 6.2 times smaller than the acceptable design variation to have greater than a 50% yield for a chip with 10^9 components.

The potential catastrophe that lies herein is that because devices are becoming so small, the variations that need to be controlled are rapidly approaching the atomic scale as demonstrated below. It simply may not be possible to produce the required level of control at an acceptable cost using the techniques we now envision. Interestingly, the consequences of large numbers was the fear of many skeptics of the IC in its early days. Attributing the discrete transistor yields obtainable in the 1950s entirely to the second term, Y_{VAR}, yield esti-

mates for ICs with component counts near n = 10 quickly approached zero. However, because the yield observed for discrete devices was actually due to the aerial term, there was no catastrophe then. The case seems very convincing that the industry will return to the situation in which device-to-device variations will be the dominant factor in determining the feasibility of further progress on monolithic integration.

The ability to cram more devices onto the IC has led to the absorption of entire systems onto a single chip, including functional blocks that traditionally were built in incompatible technologies—for example, memory, analog, and radio frequency (RF) blocks. Information processing speeds have continuously increased due to improvements in the performance of both smaller devices and new circuit or systems architectures. With the concurrent or even accelerated scaling of voltage with feature size, the IC's power efficiency (mW/MIPS) has fortunately also improved just as portable and wireless applications have become more popular and low power has become a highly important issue. Because dynamic power is proportional to CV^2 —the product of the capacitance and the square of the voltage—a function realized in 50nm/1V technology should achieve a power reduction by as much as a factor of 250 over 0.5μm/5V technology. Many products that were not considered to be portable surely will become so.

To continue achieving these improvements in the future, as well as to drive the technology to sub 100-nm dimensions and its eventual limits, a number of key problems must be solved. Independent of the device or circuit architectures employed, lithographic technology will be a primary concern because we are rapidly approaching the point at which an alternative to optical lithography—which has existed since the beginning of the IC—will be required. Many other fabrication challenges exist—for example, thin gate oxides and shallow junctions—but these are likely to evolve continuously from present capabilities until more fundamental limits of devices are approached.

Lithography

The semiconductor industry is facing a major hurdle in that conventional materials for making optical components (lenses, masks) become opaque for source wavelengths, λ, shorter than 193 nm. The ultimate resolution, L, of a lithography system can be described by the Rayleigh condition $L = k \lambda / NA$, where NA is the numerical aperture and k is a coefficient determined by process conditions. For present-day optical lithography, a standard k and good NA typically are both 0.6, so it is challenging to push resolution beyond λ—for example, L = 0.25 μm for a 248-nm deep ultraviolet (DUV) KrF source. Exposure tools using ArF 193-nm sources are expected to be introduced into manufacturing by the end of this decade but to press beyond 0.18-μm device features, improvements to NA and k would be required. One could, for instance, attempt to build higher NA optics, preserving large fields with scanning stages. However, significant improvement over today's best systems is unlikely as depth of focus scales as NA^{-2} and lens design is probably near its ultimate limits. The k factor can be improved by optical proximity correction and phase shifting masks, advanced illumination techniques, and modified resist processes. Even with these enhancements—which will involve added process complexity, lower margins, and higher costs (especially for masks)—the industry still will be hard pressed to make the 0.13-μm process generation with 193-nm lithography tools.

For these reasons, various alternative approaches to lithography are being considered. One of the oldest candidates is x-ray proximity printing, first pioneered in the 1970s by MIT and Bell Labs. In this technique, the mask must be held close to the wafer to achieve the necessary resolution—for example, for 0.13-μm resolution, the mask must be within 10 μm of the wafer. Feature sizes on the mask and wafer must be identical in size (1-to-1 or 1×), making these masks much more difficult to fabricate than the reduction masks used today. The source of x-rays most probably would be an expensive synchrotron that can supply x-rays for many writing tools simultaneously. For this reason, x-ray technology lacks granularity—the ability to add an arbitrarily small number of tools to a line at a cost proportional to the number of tools. Furthermore, it may not be economical for use in ASIC production, where mix-and-match approaches

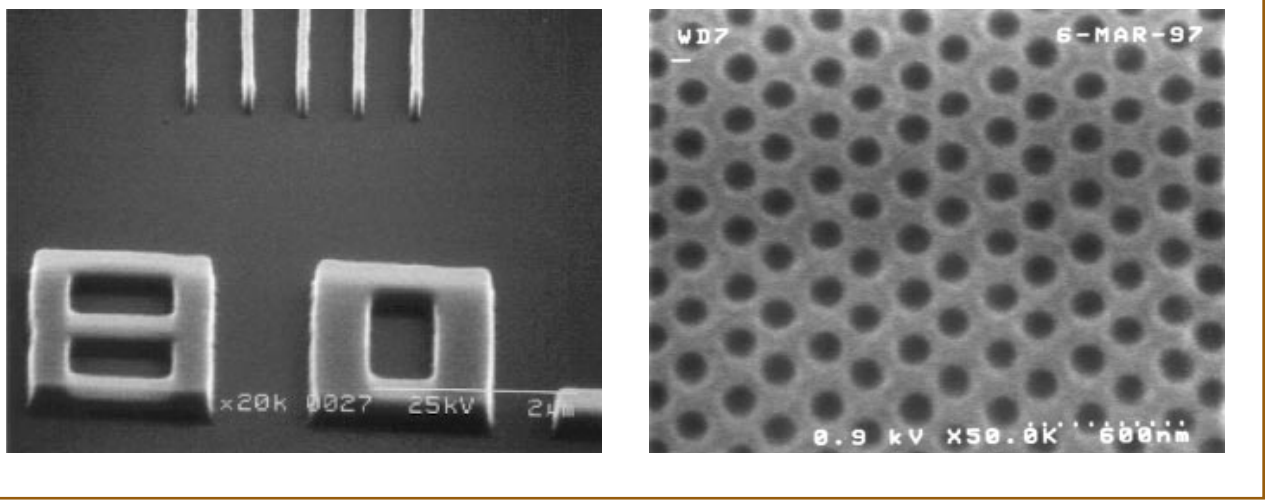

Figure 2.
80-nm features exposed using Scalpel™ lithographic technology (80-nm lines on the left and 80-nm contact windows on the right).

are frequently used. X-ray technology is also believed to be incapable of resolution much past 100 nm due to limits in fabricating 1× masks and in controlling the mask-to-wafer gap size.

A second approach is electron beam (e-beam) lithography. Because the electron has very short wavelengths, the required resolution unquestionably can be obtained. Direct-write e-beams have been used to write features down to 1 to 2 nm. However, this technique is much too slow to be used as a production wafer fabrication tool. During the past few years, Hitachi developed a projection e-beam approach called *cell projection*, which uses an absorbing stencil mask with reduction optics to produce higher throughput. With this method, a limited number of unit cell patterns (up to $5 \times 5\ \mu m^2$ in size) are printed repetitively to produce the chip patterns. In principle, cell projection could be well suited to memory chips, which contain large areas of repeating patterns. However, it suffers from limited throughput potential (one or two 8" wafers per hour) due to electron-to-electron interactions and minimums on resist sensitivities.

Recently, Bell Labs researchers Steve Berger and Murray Gibson invented an approach we call Scalpel™, which stands for SCattering with Angular Limitation Projection Electron Lithography.[4] In this approach, contrast is obtained from the difference in the angular distribution in the scattering of electrons by heavy and light atoms—for instance, by using tungsten (W) features on a thin nitride (Si_3N_4) membrane. The W atoms scatter electrons strongly so that by placing an aperture in the back focal plane of the projection system, the scattered electrons are nearly completely eliminated. In this fashion, the necessary contrast is obtained without heating the mask like an absorption system. Inexpensive magnetic lenses are used to project a reduction (for example, 4×) image, and as **Figure 2** illustrates, a resolution of 80 nm has been demonstrated. This result represents the highest resolution obtained by a projection e-beam approach.

Unlike proximity x-ray, Scalpel is a projection/reduction technology having granularity—with an overall tool footprint of approximately the same size as today's optical tools. Together with its ability to adjust magnification, Scalpel is ideal for mix-and-match applications. Even for high-throughput systems, the requirements on resist materials are not inordinately severe. These requirements seem capable of being met with conventional single-layer materials currently used for 248 and 193 nm DUV lithography.

Process latitude is another of the strong features of Scalpel technology. Because the wavelength of the 100-keV electron is so short ($\lambda = 0.0037$ nm) and the numerical aperture of the optics is so small (NA ~ 10^{-3}), the depth of focus is 100 times larger

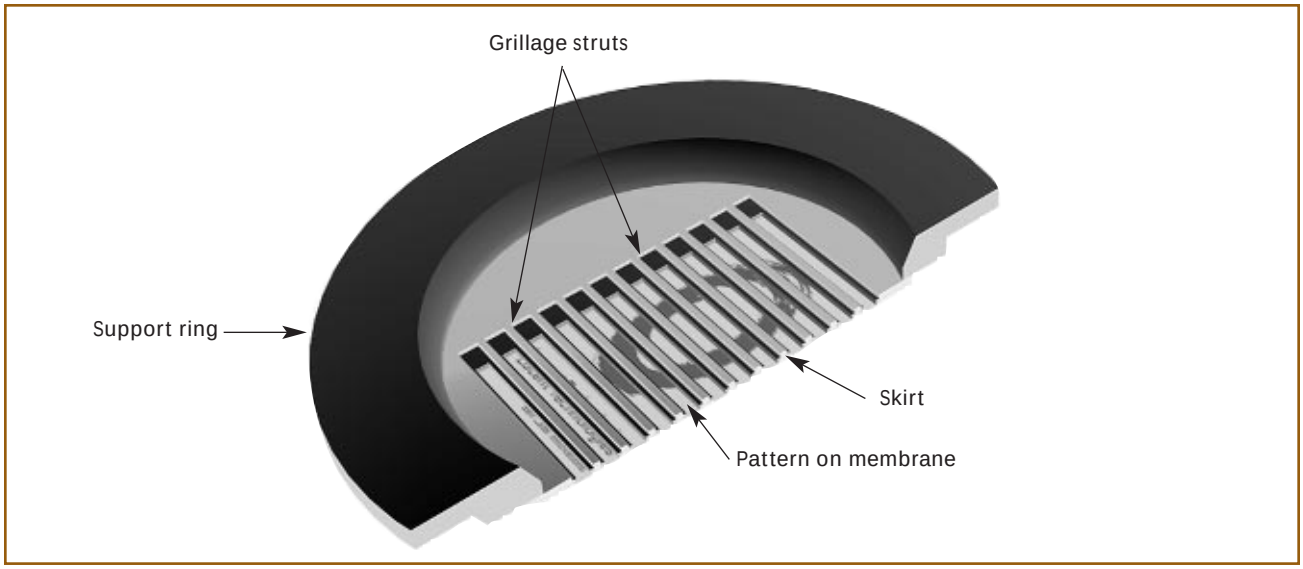

Figure 3.
Schematic of a Scalpel™ mask showing the 4× pattern written in stripes on a silicon nitride membrane and supported by a silicon wafer having a grille structure for support.

than today's optical lithography systems—on the order of 100 μm for 0.25-μm features. In addition, Scalpel affords considerable exposure latitude—nearly 30% to achieve a ±10% variation change on 0.18-μm pitch patterns over a 30-μm depth of focus.

Scalpel masks (see **Figure 3**) contain a series of struts that give it strength and control distortion. The approach to writing is to scan the mask in one direction and stitch the rows together. Such stitching has been demonstrated using cell projection and direct-write e-beam, so it should also be feasible with a Scalpel tool. With currently demonstrated technology elements, it is expected that throughput in production Scalpel systems will exceed 30 eight-inch wafers per hour. Because the exposure tool should be relatively inexpensive to produce and because 4× Scalpel masks require no sub-features, industry estimates predict that Scalpel technology should have the lowest cost per level at 0.1 μm of all known alternatives and—most interestingly—be even less expensive than 193 nm DUV at 0.13 μm. This cost advantage is accelerating Scalpel's attempt to intercept the 0.13-μm generation.

Another post-optical alternative, first demonstrated by Bell Labs some years ago,[5] is to use all reflection optics in the extreme ultraviolet (EUV) ($\lambda \approx 13$ nm) to obtain the required resolution. This approach has the advantage over proximity x-ray of being a reduction technology and, hence, may be extendible beyond 0.1 μm. However, it has several significant difficulties, including nearly perfect multilayer masks, high-reflectivity multilayers, very intense ultraviolet sources, surface-sensitive resists, and refractive optics at the limit of what is available. Work at the Sandia and Lawrence Livermore National Laboratories has resulted in many possible solutions to these problems, but numerous fundamental and practical hurdles remain. Even if a prototype can be produced, it is expected that EUV will always have a significant cost disadvantage when compared with Scalpel. Moreover, EUV is not as easily extendible because it will require proximity corrections on the masks.

Extremely high-resolution lithography has also been obtained using scanning tunneling microscopy (STM) and atomic force microscopy (AFM). Individual tunneling tips can obtain atomic resolution. However, a single STM tip is very slow, and various proposals for a "brush" of tips has been made. The number of simultaneous writing tips required for reasonable wafer throughput is somewhere between 10^6 and 10^9 depending on the exact process. Whether such an approach can ever be made reliable is a question for further research.

Transistor and CMOS Limits

Over the entire history of the IC, studies have predicted that integration limits would be reached owing to problems associated with scaling semiconductor transistors. In considering such limits, we focus here on the metal-oxide semiconductor field-effect transistor (MOSFET), which has been the driving force behind the very large scale integration (VLSI) age. Used in complementary (CMOS) configurations, the MOSFET has been predominant due to several key attributes, perhaps the most important being its ability to achieve both extremely low stand-by power—limited by leakage currents—and small, dynamic power-delay products, owing to low duty factors ($\tau_{delay} \times f_{clk}$) and the ability to use low power supply voltages. Additionally, CMOS has advantages in scalability, relative design (for example, clocked logic) and process simplicity, as well as in high functional densities.

Design strategies for scaled MOSFETs have been an intense area of investigation since the 1970s. The constant-field scaling rule[6] attempts to maintain identical field profiles (to avoid breakdown) and constant power dissipation densities (to limit self-heating) by applying a single shrink factor, κ, for all critical dimensions—for instance, gate length L, gate oxide thickness t_{ox}, and depletion widths W_D (by also scaling doping concentrations N by κ)—and $1/\kappa$ for operating voltages. In such a scenario, therefore, device currents are reduced by $1/\kappa$ while power dissipation per circuit and delay improve by $1/\kappa^2$ and $1/\kappa$, respectively. Even though this rule has not been followed exactly—for instance, modifications are required to inhibit deleterious short-channel effects[7] (see below)—it does provide a useful guideline for more accurate design procedures and for capturing general trends.

Scaling operating voltages is indeed essential to preserve reliability and control in small devices. While the industry at one time attempted to postpone power supply (V_{DD}) reductions beyond 5 V based on standards arguments, the demand for portable electronics now has accelerated voltage reduction to minimize dynamic power, as pointed out above. However, unless the MOSFET threshold voltage V_T is reduced at the same rate as V_{DD}, serious problems arise—for example, increased stand-by currents I_{off}, unacceptable noise margins, circuit speed improvement limited by power dissipation density, and increased sensitivity of the speed to inevitable variations of V_T across a chip.[8] Each of the aforementioned problems poses a fundamental limitation to CMOS.

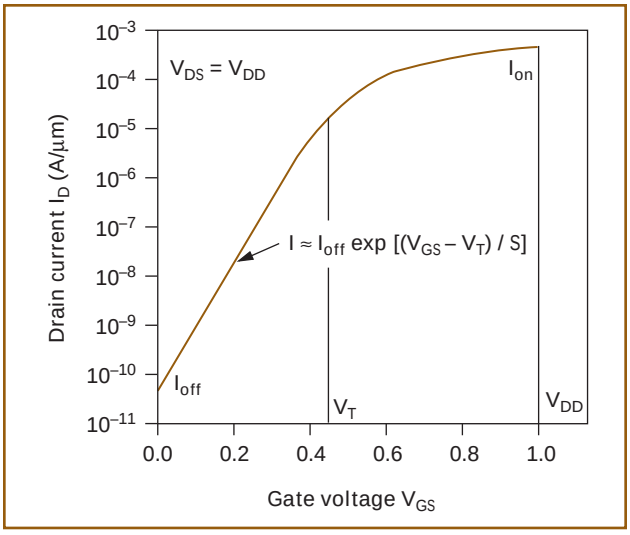

Figure 4.
Typical MOSFET subthreshold current-voltage characteristic.

To illustrate the V_T scaling problem, **Figure 4** shows a typical log drain current, I_D, versus the gate voltage V_{GS} characteristic. The on-current I_{on}, which in large part determines the speed of a CMOS logic circuit, is intimately related to I_{off}, V_T, and the subthreshold swing S, which depends on ambient temperature and device design. A key metal-oxide semiconductor (MOS) device design specification is the acceptable value of I_{off}, bounded either by stand-by power dissipation through idle CMOS gates ($P_{static} = n_{gate} I_{off} V_{DD}$, where n_{gate} increases as κ^2) or by hold time requirements on dynamic nodes—for instance, in DRAM cells. In contrast, one would like to maximize I_{on} to minimize gate delays, so the ratio I_{on}/I_{off} is a key measure of the quality of the transistor switch. This analysis suggests designing MOSFETs with S as low as possible—for instance, by increasing the oxide capacitance to depletion capacitance ratio. However, the minimum value of S is limited to kT/q per e-fold change in current, or about 60 mV per I_D decade at room temperature.

The constant-field scaling rules were generated from one-dimensional models of the MOSFET, valid in

the long-channel regime in which the gate length is substantially larger than the depletion widths induced at the source-drain junctions. However, as MOSFETs approached 1-μm gate lengths, these models were no longer adequate. Such effects as V_T roll-off with decreasing L and elevated drain bias (called drain-induced barrier lowering [DIBL]) have forced device designers to consider two-dimensional effects on key characteristics, like V_T and S. Control of these effects typically requires higher doping in the channel and scaling perhaps as κ^2 instead of the constant-field rule κ. However, the doping cannot be uniform, otherwise V_T will not scale correctly. As a result, a key part of the design of submicron MOSFETs has been engineering the doping profiles using sophisticated numerical process and device simulation programs. Even with near-optimal profiles, other MOSFET parameters must be compromised to limit such effects as V_T roll-off, including increasing gate capacitance, body coefficient, and—most unfortunately—the subthreshold swing S. In fact, the required increase in doping is expected to be such that S ≈ 70-80 mV/decade may be a best-case assumption for room temperature operation of sub 100-nm CMOS devices in bulk silicon substrates.

The linear dependence of S on temperature might suggest that much better performance could be achieved using cryogenic cooling. The progress in micro-refrigeration has not motivated the industry to take this direction, however, not even for rack-mounted systems. Refrigeration is not viable at all for portable, power-limited, or cost-sensitive applications that make up the majority of the IC market. Therefore, even with perfect MOSFET design, a minimum limit must be imposed on V_T for conventional CMOS—probably on the order of 0.4 to 0.5 V at 50-nm gate lengths—thereby also limiting the minimum V_{DD} to 0.8 to 1.2 V based on noise margin considerations. In turn, these voltages will impose limits on physical dimensions due to breakdown and reliability thresholds.

With other device parameters properly scaled (for example, source/drain resistance and overlap capacitance), optimal performance can be achieved by maximizing the gate-to-channel capacitance, C_G, of the MOSFET—for instance, by minimizing the associated oxide thickness t_{ox}. High C_G maximizes I_{on}, minimizes S, permits acceptably low V_T, and gives good short-channel control. However, t_{ox} cannot be reduced without limit, especially when V_{DD} can no longer be reduced due to gate leakage associated with the high operating field, $\approx V_{DD}/t_{ox}$. For oxides below 3 nm, gate leakage becomes dominated by direct tunneling current, which must be limited to meet both stand-by power specifications (similar to I_{off} above) and reliability requirements, which tend to be related to the integrated current put through the oxide over its lifetime.[9]

Theoretically, it would seem straightforward to achieve the minimum t_{ox} associated with the maximum tolerable gate leakage. However, specific process details will determine how close one can come to the ideal t_{ox}, and hence C_G. Of particular importance is the relative smoothness of both the gate-to-oxide and oxide-to-substrate interfaces because microscopic roughness exponentially enhances the tunneling current through both local thickness minima and electric field maxima from sharp corners (see **Figure 5**). Clean surfaces and low-defect material are also critical. Further, even a perfect as-grown gate oxide can be degraded by plasma-induced damage in the back end of the process in which high-temperature anneals are precluded.[10] Assuming the plasma damage problem can be controlled, the minimum t_{ox} is expected to be in the range of 1.2 to 2.0 nm.

Further gate capacitance improvements might be possible by integrating a higher dielectric constant material—for example, Ta_2O_5 or Si_3N_4—into the gate dielectric. In principal, these materials would allow a proportionally higher voltage at an equivalent capacitance or a higher capacitance for the same voltage. However, it has been problematic to integrate these materials due to such issues as interface states, leakage, and high-temperature incompatibilities. Promising results have been achieved by putting layers in a sandwich between two thin SiO_2 layers,[11] and this approach may be attractive for maintaining compatibility with higher supply voltages. However, when factoring in manufacturing variations (see below) in all the layers, it is unlikely that oxide sandwich structures would offer a practical advantage in extending the ultimate limits of C_G.

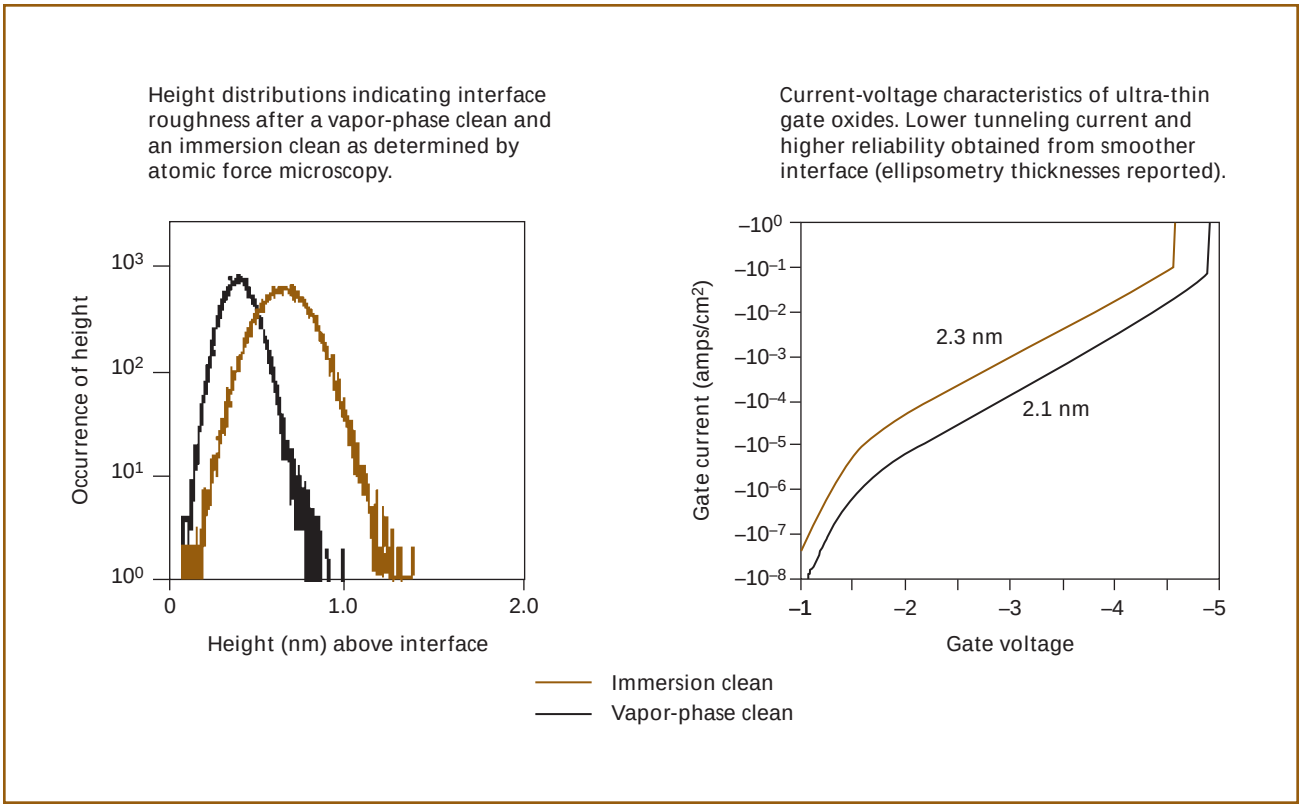

Figure 5.
Graphs showing lower tunneling current resulting from a smoother Si/SiO_2 interface.[12]

In determining the achievable limits of gate capacitors, other important factors must be considered. For instance, the actual effective electrical gate capacitance is known to be lower than the value implied simply by the intrinsic dielectric (for instance, ε_{ox}/t_{ox}) due to depletion effects in the gate material and quantization in the channel.[13] Minimizing the effect of quantization in any substantive way requires a major change in device structure away from the ordinary MOSFET. Gate electrode depletion can be avoided by heavily doping the respective polysilicon gates, but great care must be taken to get the doping to the interface, keep it activated, and not let it cross the oxide and compensate the channel. The latter constraint is especially difficult for low V_T positive MOSFETs (PMOSFETs), and nitrogen engineering of the predominantly SiO_2 film seems like a promising approach to minimizing this effect.[14] (Relatively small amounts of nitrogen may have other beneficial effects—for example, better reliability.) Metal gates would be another solution to the gate depletion problem, but severe integration problems and undesirable (typically midgap) electron affinities are enormous barriers. Metal and/or silicide straps on top of a polysilicon gate are essential for minimizing resistance and for effectively strapping respective n-type (NMOS) and p-type (PMOS) MOS transistors.

Scaling the MOSFET toward its limit also involves scaling the source and drain junctions. If V_{DD} cannot be scaled indefinitely, another possible leakage constraint is direct tunneling from the drain junction to the substrate, exaggerated by the high levels of doping that might be used. Using simple device structures, gate leakage looks to become a first-order concern at gate lengths near 50 nm, although carefully optimized doping profiles may help manage this problem. Another key concern with source/drain engineering involves concurrently minimizing junction depth and series resistance.[15] In particular, transient-enhanced diffusion effects impose limits on ultra-shallow pro-

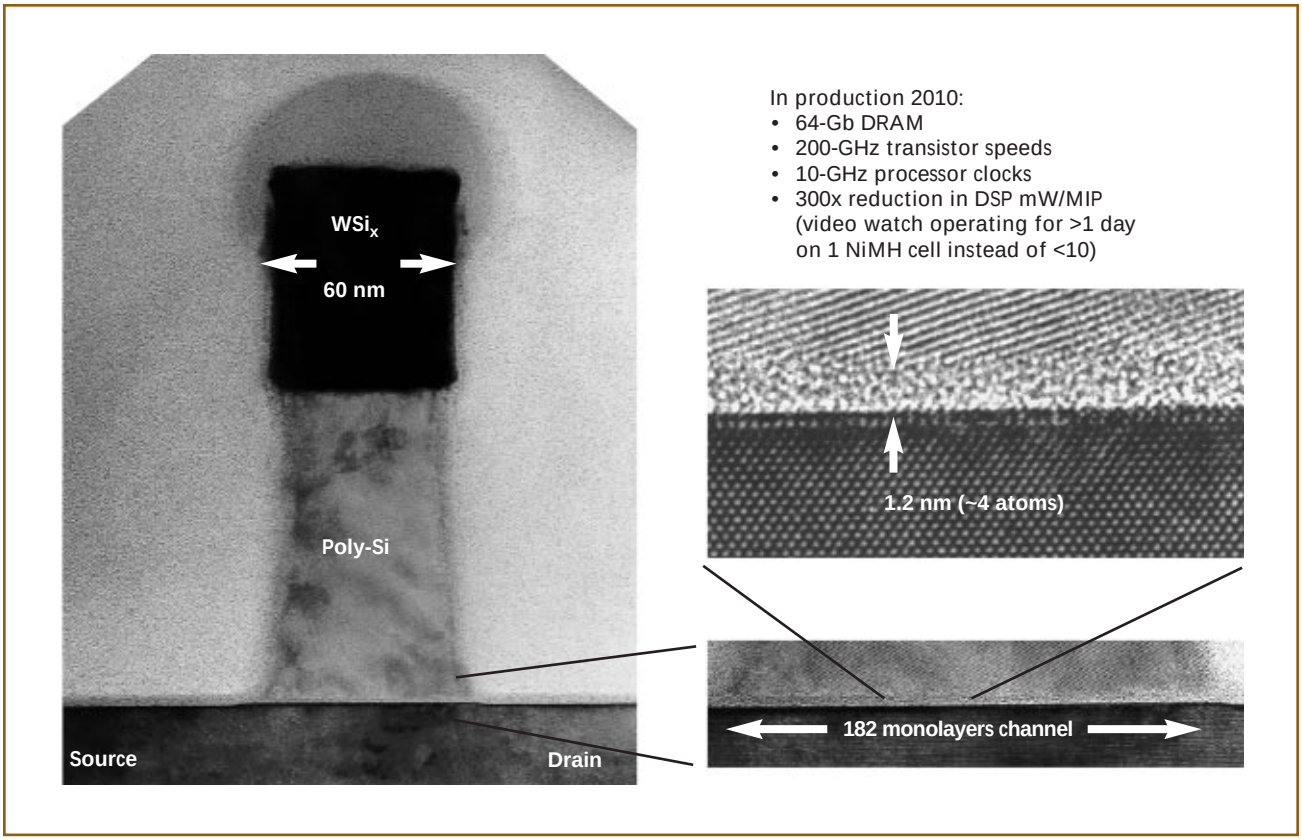

Figure 6.
Cross-sectional micrographs of a 60-nm MOSFET built at Bell Labs with the atomic lattice visible in highest magnification view of 1.2-nm gate oxide.

files, even when minimal implant energies are used. If direct implantation is not adequate, other solutions that might be adopted include raised source/drains or other diffusion sources. In any event, junction doping cannot be increased arbitrarily because solid solubility limits begin to encroach at impurity concentrations $>10^{20}$ cm^{-3}, so there might be a legitimate concern that junction series resistance could eventually limit MOSFET performance. Finally, while problems associated with substrate current and hot carrier reliability might be expected to disappear at minimum V_{DD}, the sharp electric fields and shallow junctions in ultra-small devices will amplify ionization feedback effects, which could still stimulate high-energy carriers[16] and likely require new engineering solutions.

Several research groups have demonstrated the feasibility of CMOS to <100 nm (**Figure 6**).[17] Using ultra-fine e-beams or resist ashing to achieve the requisite dimensions—albeit not scaleable over full wafers—and oxides as thin as 1.2 nm, stunning performance has been achieved. These results include n-channel MOSFETs with cutoff frequencies of 120 GHz and CMOS circuits with gate delays of ~10 ps, all at room temperature.[18] More recent dc results on 60-nm gate length devices demonstrate the largest drive current, I_{on} ~ 1.8 mA/mm, and transconductance, g_m ~ 1.12 S/mm, ever reported for a MOSFET.[17]

Feasibility of individual ~ 50-nm devices and small circuits, while an important and impressive result, unfortunately does not prove full viability of the Moore Plot to the 50-nm generation—estimated by the 1997 Semiconductor Industry Association (SIA) roadmap to begin production in 2012. As discussed earlier, the magnitude of the effects of manufacturing variations and process statistics on device and circuit performance may be such that the probability of making a working IC with the associated number of components (256G DRAM, logic chips with 1.4G transistors) effectively could be zero. MOSFET characteristics are affected by three major sources of process

variation: gate oxide thickness, gate length (via lithography and etching), and the impurity profile (from implantation and diffusion). Without a major change in processing techniques, it will become ever more difficult to control each of these properties across chips with billions of transistors.

For instance, a 50 × 50-nm transistor translates to about 150 silicon atoms in gate length (135 atom-wide channels) and, for a well designed transistor targeted for 1-V operation with V_T = 0.45 V and t_{ox} = 1.5 nm, only about 175 dopant atoms in the channel depletion region. Numerical simulations using PADRE[19] predict that a 10% variation in each t_{ox} (± 0.15 nm or one lattice step), channel impurity profile (17 dopant atoms), and L (± 5 nm or 16 silicon atoms) corresponds to between ± 20-40 mV in V_T or about ± 60 mV, taking the effects of all three together. Similarly, I_{on} (hence, logic gate delay) thereby varies ± 20-30% and the maximum I_{off} is 15× nominal, worse than the change implied by just ΔV_T due to DIBL effects. Even without DIBL, a normal distribution in V_T with σ_{VT} = 60 mV increases stand-by power by almost 40% over the nominal $I_{off} \cdot V_{DD}$.

Synchronous circuit architectures, which by far make up most of the ICs today, depend on a global clock signal whose speed is determined by the worst-case gate delay. Therefore, the 10% process variation for the device above could slow the clock and, hence, the overall speed of the system by as much as 30%. This situation should be compared to today's 0.5-μm technology, where the speed variation is typically 10%. As long as the threshold cannot be scaled, this penalty will get worse without significant improvements in process control. The situation is even worse for mixed-signal ICs, as many analog circuit blocks depend on precise transistor matching.

Hence, while a hard limit to CMOS technology cannot be extracted, a number of serious challenges have been described, led by the difficulty in continuing to scale voltages while maintaining low stand-by currents. Thus, the 50-nm generation seems to be a best approximation to the point where these challenges to extrapolations of current implementations are serious. In the following section, we examine some modifications to CMOS technology that may help to extend these limits and to develop more revolutionary alternatives.

CMOS Modifications and Possible Alternatives

Will there be a replacement to CMOS for VLSI applications? Because of the enormous learning base behind silicon CMOS, such a change is unlikely on a broad scale, at least until the absolute limits of the technology are reached. Extending the CMOS architectures of today, the most likely roadblock is the difficulty in scaling V_T to obtain adequate design margin while keeping I_{off} and stand-by power within acceptable bounds—exaggerated by the required degree of statistical control. Before examining more revolutionary alternatives—attempting to address general-purpose or niche needs—it is first instructive to look at the evolutionary changes that might be made to conventional CMOS, with such changes focusing on the V_T and I_{off} scaling dilemma. Interest in these ideas has been heightened by the advantages in power dissipation to be gained by introducing low V_{DD} processes at earlier generations.

New process methods for conventional MOSFETs may provide more precise control over feature definition, film thickness, and impurity placement. For instance, molecular beam epitaxy might be used to achieve delta doping in the MOSFET channel, thereby improving the low V_T and DIBL design tradeoff while also minimizing the effect of dopant fluctuations. It might also be possible to develop new approaches to designing the MOSFET that could reduce process sensitivity. While these types of proposals seem to be a promising avenue of investigation, they probably offer only incremental improvements without a more dramatic change to the device architecture. Furthermore, as emphasized in the introduction, the cost of such solutions must not overwhelm the advantage of further integration or they will not be pursued.

Another potential solution—already touched on—is to attempt to improve the MOSFET subthreshold swing S, thereby allowing V_T to approach zero as closely as possible while maintaining an acceptable I_{off}. MOSFETs on silicon-on-insulator (SOI) substrates offer an incremental advantage over those produced on bulk wafers. By using a very thin silicon

film—thinner than the gate depletion depth in a bulk device—almost ideal thermally limited S can be achieved. Additionally, very thin SOI films eliminate junction area capacitance and leakage and simplify device isolation. Numerous problems need to be overcome with SOI MOSFETs, including floating-body effects, self-heating, higher substrate costs and defect densities, and film thickness sensitivity/control issues, which can directly affect both V_T and the quality of ultra-thin gate oxides.

A promising path of CMOS evolution is toward the use of varying device thresholds—for instance, by using modifying local channel implants or oxide thickness to produce different static V_T values. A high V_T could be chosen for memory access devices (requiring low I_{off}), while a lower V_T would yield acceptably higher I_{on} for dynamic logic. Because memory consumes a large portion of the device count—even for ASIC and logic applications—the potentially high I_{off} of the logic transistors may not be the determining factor in stand-by power. Furthermore, new logic circuit architectures are being investigated that use different logic V_T levels to achieve lower stand-by power and high current drive.

An even more attractive possibility would be to vary V_T dynamically under on-chip bias control. One such proposal uses a low nominal threshold V_T^0 to obtain high-speed logic, then selectively shuts off portions of the chip by applying negative V_{BS} (positive for PMOS) to transistors in isolated wells.[20] For the 50-nm MOSFET mentioned above, a 200-mV upward shift in V_T —corresponding to a factor of ~10^3 in I_{off} — would result from applying –1 V to the substrate. However, this approach is limited by constraints on junction leakage, breakdown, and capacitance, as well as on the difficulty in simultaneously achieving low V_T^0 and S, yet high V_{BS} coupling. From a circuit point of view, an extra voltage must be generated, and the latency in shutdown makes it likely that this approach can only be used for sleep mode and not in the midst of normal chip operation.

SOI substrates offer the unique ability to contact separately the substrate of individual MOSFETs, leading to several advantages over bulk CMOS for dynamic V_T control. For instance, substrates of individual transistors of both types can be contacted (albeit at a nominal cost in process complexity and area), and no well junction capacitance effectively exists. Further, by designing NMOSFETs with *high* V_T^0 and, hence, low nominal I_{off}, a *positive* V_{BS} (negative for PMOS) can shift the V_T downward to a high current drive mode.[21] These structures require neither an extra voltage level nor the aggressively thinned silicon layers that fully depleted SOI calls for. Again, using the 50-nm device design above, substrate-to-gate strapping achieves ≈13-mV/decade improvement in S and ≈50% improvement in drive current at V_{DD} = 1V (V_{BS} = 0.6V) for the same nominal I_{off}. A limitation of this technique is that the positive excursion of V_{BS} must be limited so that the well-to-source junctions do not become forward biased. To maintain a logic swing above ≈ 0.6 V, at least one extra transistor is required in each logic input.

Another interesting proposal is the dual-gate MOSFET, which provides both a ground plane to suppress DIBL[22] without compromising junction leakage and a means of adjusting V_T dynamically without junction forward bias.[23] Additionally, these structures may alleviate dopant fluctuation but trade this problem for channel thickness control. Other significant challenges to realization include contact to the back gate, oxide quality (especially in a buried back gate), series resistance to the ultra-thin channel, as well as back gate alignment, parasitic capacitance, and work function. While this is an interesting avenue of research, a convincing demonstration of these structures has not yet been given and manufacturability is likely to be a persistent question.

Turning to more dramatically different alternatives to CMOS, many have been proposed over the years and most have been inspired by attempts to gain better electrical performance (higher speed) rather than attacking the perceived limits to integration—that is, the looming I_{on}-I_{off} crisis. Additionally, new structures more often than not address only one segment of needs—for instance, a denser memory cell as opposed to general-purpose logic. Because it is impossible to cover all potential alternatives in every area relevant to electronics, the following brief discussion focuses on important underlying themes and a limited number of

concepts that could evolve into the most relevant and widespread applications.

New device concepts might evolve from a number of different directions. For instance, the modifications of CMOS described above involve using known materials and physical principles and assembling improved structures, generally using the extra degrees of freedom available geometrically. Another realm of investigation attempts to apply different physical effects—typically originating from newly developed materials, processes, or solid-state phenomena appearing due to reduced dimensions—to produce new *and useful* electrical components. It is the "useful" part of this statement, however, that has been a hurdle. Researchers have pursued these avenues since the discovery of the transistor (for example, tunnel diodes and superconducting circuits). Yet, it is difficult to point to even a few major commercial successes, a situation affected in part by the continued evolution of silicon MOS technology. On careful examination, most ideas to date have exhibited severe drawbacks in a highly integrated system. However, they do represent the frontier of understanding. Only by recognizing the areas in which the true limits of CMOS might exist, by targeting relevant goals, by fully understanding system constraints, and by dropping fruitless directions will promising new solutions emerge and be further developed.

A successful example, emanating directly from the MOSFET itself, is the concept of nonvolatile memory. It was recognized quite early on that hot carriers could be generated at elevated voltages, and these carriers could be used to charge a floating gate. This concept led to the development of erasable programmable read-only memory (EPROM), electrically erasable programmable read-only memory (EEPROM), and flash memory products. As dimensions have continued to be scaled, hot-carrier effects have become pervasive and new mechanisms of energy exchange have become relevant. For instance, using ionization feedback effects, nonvolatile memory cells have been built that operate on voltages lower than the Si/SiO_2 conduction band,[24] something previously thought to be impossible. Memories built using this concept are very amenable to scaling and have advantages with respect to speed and power. Further, because they achieve excellent intrinsic control of the floating gate charge, they are a very promising means to achieve multibit storage—that is, gaining density improvements by packing more information in a single device.

Along a similar line, hot-carrier effects and thermeonic emission can be realized in semiconductor heterostructures—an effect known as *real space transfer*[25]—and then employed to make new classes of devices. One such example is the charge injection transistor (CHINT), whose negative resistance characteristics and intrinsic symmetry have been used to implement a single device that—with a load—can implement logic operations requiring up to five times more transistors in CMOS.[26] These types of structures have been described as functional devices because they intrinsically perform much more complex tasks than just amplification or switching. Interestingly, the concept of functional devices has been suggested for more than thirty years as a means to overcome the limits of ICs both in the difficulty of large-scale manufacture and in the complexity of design.[27]

Another class of devices has evolved out of the quantization effects associated with finer and finer feature sizes. Most proposals to date center around an island region separated by energy barriers from source and drain regions—for instance, resonant tunneling transistors (RTTs) and single-electron transistors (SETs). In the RTT,[28] the gate bias changes the potential of the island to bring it into resonance, allowing tunneling carriers to transfer from source to drain. SETs typically work via the principal of the Coulomb-blockade,[29] whereby changes in the applied gate voltage corresponding to a single electron can switch the source-to-drain current on. In the off state, current is precluded by electrostatic repulsion. RTTs and SETs have been configured as functional devices and as memories.

SETs are often discussed as an eventual successor to the MOSFET, although the MOSFET itself would eventually become a single-electron device somewhere below 10 nm. Several groups have built devices using small islands of charge, both as a traditional switching device and as a memory element. While initial feasibility has been demonstrated, device and circuit architec-

tures still require substantial research to overcome severe problems, such as thermal fluctuations, process variations, sensitivities to background charge, and noise. Other difficulties include the tradeoff between low leakage (large barrier) and high drive current (short transfer time), charging long bit lines, and such parasitic effects as disturbs. It is not clear that SETs will ever work well enough at room temperature.

New materials and process technologies underlie all device advances, and in some cases, interest in promising materials and processes might be the source of device innovations, as was the case for semiconductors and the transistor. Such expectations (certainly premature) were indeed raised for superconducting materials but thus far, they have not been fulfilled. Ferroelectrics, which can modulate substrate conductivity through polarization, have also long been investigated. Although significant progress has been made in the use of ferroelectrics in nonvolatile memory (fast low-voltage switching, good retention, and endurance), serious problems still remain with integration that may stymie widespread use.

Electronic applications of organic materials and molecular processing are a more recent field of research, primarily motivated by the potential for more cost-effective fabrication methods. Impressive advances have been made on organic analogues to semiconductor transistors.[30] While these always are likely to suffer from inferior performance compared to silicon, they might be adopted for some key applications through advantages in cost (but only on large transistor size scales) and mechanical properties—for instance, for low-cost, light-weight, flexible displays. Molecular electronics, where wires and switches are made of individual or small groups of molecules, are even more speculative at present but perhaps hold even greater promise. By attempting to employ chemical synthesis to produce and self-assemble billions of identical nanometer-scale structures, the hope is to develop a technology that would be significantly cheaper per function and would not be limited by process variations. In addition, as IC interconnect—the subject of the next section—may be a key constraint, the density of connections in biological systems like the human cortex might suggest that molecular electronics could have other important advantages. It should be recognized, however, that the area of molecular electronics will be, at best, a complement to semiconductors, and it has a long way to go to even establish itself as something more than an academic research endeavor.

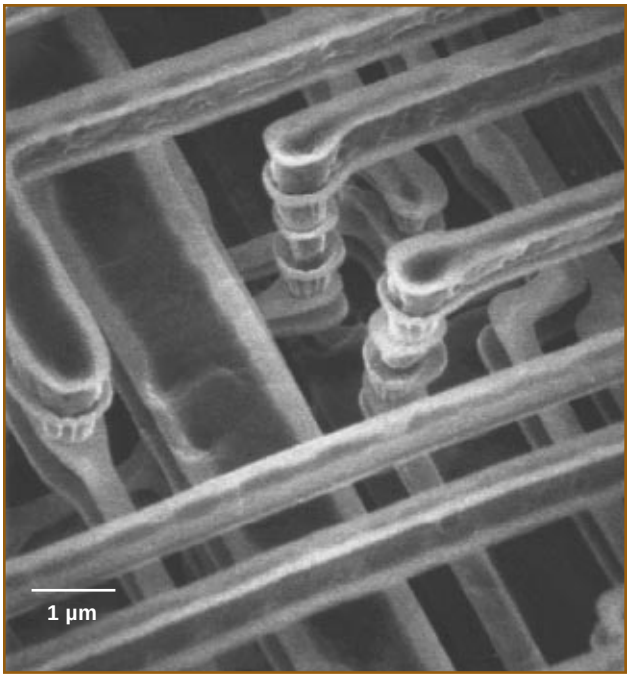

Figure 7.
Multilevel interconnect in 0.35-µm CMOS technology showing planarized levels, dense and wide pitch metal, and stacked via connections.

Interconnection

With the increasing complexities of designs and larger chip sizes, interconnect technology has evolved into a primary concern. To address the issues concerning density, planarized multilevel metal processes have been developed, which permit fine pitches on any level and dense via connections between levels (see **Figure 7**). At the penalty of adding more masks and complex steps to the IC process, multilevel metal has eased the task of hierarchically designing chips from macroblocks to large cores using local (between gates) through to global (between larger cores) levels of interconnect.

However, as feature sizes have scaled to below 0.5 µm, the performance of large chips has become more and more dominated by interconnect parasitics. The

problem is straightforward to understand by examining the scaling of RC delays, but it is essential to examine the effects for both local and global lines. For dense lines of length L, $RC \propto L^2/ws$, where w is the width of the line and s is the spacing between adjacent lines. Using a scaling rule that reduces all dimensions by the factor $\kappa \sim \sqrt{2}$, the RC delay for local interconnect stays constant while global RC increases as $\kappa^2 \sim 2$. For 0.25-μm aluminum lines and an SiO_2 dielectric, the RC delay for a minimum-pitch 1-cm line is ~10 ns, which would limit clock speeds to <100 MHz—well below the 750 MHz or so that one would expect from such a technology. Thus, global interconnect for large chips becomes a serious constraint.

A solution that has been used is to increase the metal pitch for longer interconnect on upper metal layers so that the charging time does not approach the clock frequency of the circuit. For example, if the metal layer for the 0.25-μm technology could have 1.0-μm lines and spaces, the RC delay for a 1-cm line would be ~0.6 ns, or less than half the period of a 750-MHz clock. However, continuing to extend this approach will become more difficult as increasingly smaller transistors imply faster clock speeds (hence, ever smaller RC) and longer lines across larger chips. The capacitance of very fat lines eventually will be dominated by the coupling between vertical levels of metal (rather than adjacent lines on the same level) due to the difficulty in making arbitrarily deep vias. Dynamic power increases due to the increasing wiring capacitance. Further, and perhaps most importantly, the wider the metal pitch used for routing the larger the number of metal layers that will be required, perhaps accelerating to the point where the process cost (increased steps and lower yield) make this approach unattractive.

The onset of RC versus the number of metal levels has typically occurred first in large high-speed microprocessors and is exacerbated by the partitioning of the overall design. It should be stressed, however, that the above analysis is oversimplified somewhat. In calculating signal path delays, one also needs to take into account gate switching delays, the resistance of the driver to the metal line, propagation times, and the input capacitance of the stage following the metal line. In particular, even for the long lines in today's technologies, the delay comes primarily from the driver and the capacitance of the line, not the resistance of the interconnect itself. Under these circumstances, one would not want to fatten the wires, and wide transistors (or BiCMOS) can be very effectively used to help improve performance.

One must also carefully consider electromigration (EM) effects, which put a limit on the current density in the wires and via connections between levels. A typical EM current limit rule is to stay below 2×10^5 A/cm^2 for aluminum lines. This problem has not been so severe due to the use of metal stacks with cladding shunt layers, bamboo or near-bamboo aluminum lines that inhibit mass flow, Al-Cu alloys, and the effect of cycled versus DC stress. The current density crisis can also be avoided by not scaling the height of the metal line, which is possible using the so-called damascene process. However, a major concern could be the EM effects in very small vias that violate the current density criterion already at the 0.35-μm sizes. Cladding layers in the vias again have helped avoid a crisis, but this problem bears continued attention.

Returning to the problem of performance and wiring RC delay, various solutions have been proposed. Low dielectric constant (low-k) insulating layers—such as organics, aerogels, or even air—could be used to reduce capacitance by a factor of 2 or more versus SiO_2. In addition to reducing parasitic delay, this can also improve dynamic power. The use of copper as a substitute for aluminum is attractive due to a factor of 2 lower wire resistance. Copper also has advantages in increased EM tolerance and better compatibility with the low-k insulators. Numerous difficulties affect the processing of copper, though—particularly the need for special liners (typically of high resistance) to prevent it from diffusing out of the metal lines. It should be stressed, however, that introducing low-k materials will only have a substantive effect if critical delays are not dominated by input capacitances of sequential gates, while copper will only be effective in cases in which wire resistance is a dominant concern. In cases in which both interconnect R and C have a major effect on performance, the factor of 2 to 4 in wire RC reduction still may not have an overwhelming effect. However, in such cases, it does help

to improve the tradeoff of density versus the required number of metal levels as long as the cost of copper/low-k processing is not substantially higher than the cost of aluminum processing.

Beyond metals and air, there is little one can do to improve interconnect performance from a materials point of view. For instance, one could use optical interconnects for selected signals, like clock distribution. Bell Labs has developed a technology for fabricating a large number of detectors and modulators onto silicon chips that could allow the creation of a global clock. Such a solution would lead to a major change in the manner in which we design electronics, but it may become practical because of other applications that transmit large blocks of optical data. Other speculative solutions include on-chip RF transmitters and receivers or superconducting wiring—the latter suffering from both the need for refrigeration and from the lack of adequate current carrying capability.

A more likely solution to the potential interconnect bottleneck is that the design of chips will change, both in more optimal use of the multiple metal levels and in new architectures that avoid these limitations. A higher-level design methodology will be required that combines the physical design of the devices with the logic design and placement of the individual transistors and circuit blocks, as well as the geometry of the interconnect. These will all need to be optimized with respect to propagation delay, signal strength, noise, and crosstalk.

Design—an Extension of Limits or the Ultimate Barrier?

As the basis for the analysis of CMOS limits, likely constraints on processes and devices were developed from traditional design requirements. However, as with the conclusions from the analysis of interconnect, it is possible that new circuit design techniques and architectures might help extend the limits of IC technology. While the specifics of this topic are not the primary thrust of this paper, a strong case can be made that solutions to the eventual scaling challenges are more likely to be found through coordinated efforts in CMOS process technology evolution *and* design innovations rather than through revolutionary device replacement, which also requires new circuits.

Asynchronous logic[31] offers the possibility of relieving one of the most significant burdens in interconnect technology by eliminating the need for high-speed clock distribution. Dynamic power is also reduced by not having the CV^2f from the clock. Further, asynchronous systems tend to be less affected by component variations because typically, the overall system delay is proportional to that of the average data path rather than to the worst case. As with a number of the other possibilities mentioned here, asynchronous logic is significantly more difficult to design and also can incur added circuit overhead.

The incorporation of more fault-tolerant and self-calibrating circuits can also help cope with the potentially much larger variation of device characteristics. Simple circuits, which require extremely precise parameter matching, may have to be modified substantially or avoided altogether. Circuit innovations combined with technology changes, including the incorporation of multiple threshold voltages, can suggest improved circuit blocks,[32] BiCMOS circuits for driving loads, and inductors for lower voltage CMOS RF applications. Dynamic V_T circuits offer a host of new opportunities for circuit innovation, not the least of which is the addition of more chances for local self-calibration.

In addition to asynchronous logic, a variety of other possible design-driven solutions can be used to solve the interconnect problem, including better hierarchical design systems (especially those that tend to minimize interconnect length), shared bus architectures, and better automated use of drivers, repeaters, and other techniques to reduce transmission line delay and skew. In the category of much more exploratory proposals, multilevel logic would be included that, in theory, could reduce total wiring and signal swings. Multilevel storage has been demonstrated by several organizations as a realistic solution, at least for nonvolatile memory.

Even aside from the more remote solutions, a very important caution must be raised regarding the argument that design innovation holds the key to extending integration limits. Most such solutions—for example, asynchronous logic or self-calibrating cir-

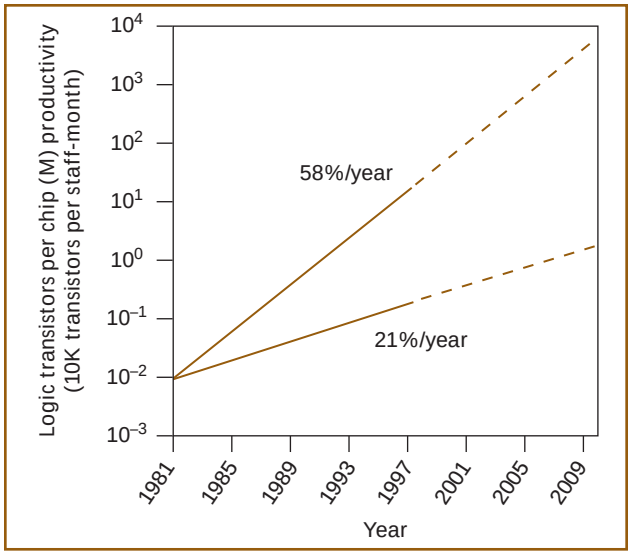

Figure 8.
Comparison of integration capability provided by process technologies and designer productivity, normalized to coincide in 1981.

cuits—required additional silicon area for implementation. Most importantly, however, almost all proposals of this nature will further complicate CAD tools and the designer's job. Even today, while still not approaching the limits to the technology, the industry faces a major design resource crisis.

A recent analysis performed at SEMATECH (see **Figure 8**) predicts a widening gap in the capability of process technology and the productivity of designers. Even with progress in CAD systems, the data show that the number of transistors available has outstripped the ability to design at almost a 3-to-1 rate, implying that more design resources are required at every generation to take full advantage of continuous improvements in process technology. The analysis implies that a given product roadmap will necessitate an increasingly larger number of designers, as well as higher costs. Further, a clear path has not even been established to fully verify and test the massive circuits of the future.

The topic of design complexity and future trends is examined in greater detail in the paper by Dunlop, Evans, and Rigge.[33] In concluding this paper, we leave open the question of whether design will help extend the technology's limits or whether complexity will be its ultimate bottleneck.

Summary

In fifty years, the world has advanced from the invention of the first transistor to a $150 billion industry that produces thousands of transistors per day for every human being on earth. Continued technical progress, at an astounding rate comparable to that of the last several decades, seems likely through at least 2015, with a principal challenge being the transition from optical lithography to a higher resolution alternative. Fundamental limits—arising from the interrelationship among physics, economics, and complexity—are not likely to surface until after 2010, when MOSFET channel lengths will be on the order of 50 nm, or less than 200 atoms long.

The most likely extensions to the technology will come through relatively minor modifications to the transistor structure—perhaps, for instance, dynamic threshold control—but hopefully, through much better design and testing techniques. The likelihood of the development of an outright replacement to silicon IC technology is small, but various avenues of research could potentially lead to complementary solutions for certain classes of applications.

Acknowledgments

The authors gratefully acknowledge the helpful discussions with R. C. Liu, G. L. Timp, and B. E. Weir of Bell Labs. They also thank Bell Labs researchers F. H. Baumann for providing the TEM images used in the figures, and R. M. Camarda for furnishing the schematic of the Scalpel mask.

References

1. R. W. Keyes, "Physical Limits in Digital Electronics," *Proceedings of the IEEE*, Vol. 63, No. 5, May 1975, pp. 740–767.
2. G. E. Moore, "Intel—Memories and the Microprocessor," Reprinted with permission of *Daedalus, Journal of the American Academy of Arts and Sciences*, from the issue titled "Managing Innovation," Spring 1996, Vol. 125, No. 2, pp. 55–80.
3. J. T. Clemens, "Silicon Microelectronics Technology," *Bell Labs Technical Journal*, Vol. 2, No. 4, Autumn 1997, pp. 76–102.
4. J. A. Liddle, L. R. Harriott, and W. K. Waskiewicz, "Projection Electron Beam Lithography: Scalpel," *Microlithography World*, Vol. 6, No. 2, Spring 1997, pp. 15–18.
5. J. E. Bjorkholm, J. Bokor, L. Eichner,

R. R. Freeman, J. Gregus, T. E. Jewell, W. M. Mansfield, A. A. MacDowell, E. L. Raab, W. T. Silfvast, L. H. Szeto, D. M. Tennant, W. K. Waskiewicz, D. L. White, D. L. Windt, O. R. Wood II, and J. H. Bruning, "Reduction Imaging at 14 nm Using Multilayer-Coated Optics: Printing of Features Smaller than 0.1 μm," *Journal of Vacuum Science and Technology B*, Vol. 8, No. 6, Nov.-Dec. 1990, pp. 1509–1513.

6. R. H. Dennard, F. H. Gaensslen, H.-N. Yu, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of Ion-Implanted MOSFETs with Very Small Physical Dimensions," *IEEE Journal of Solid-State Circuits*, Vol. SC-9, No. 5, Oct. 1974, pp. 256–268.
7. J. R. Brews, K. K. Ng, and R. K. Watts, "The Submicrometer Silicon MOSFET," *Submicron Integrated Circuits*, edited by R. K. Watts, Wiley-Interscience, Chichester, UK, 1989, pp. 9–86.
8. G. Timp, R. E. Howard, and P. M. Mankiewich, "Nanoelectronics for Advanced Computation and Communication," *Nano-Science and Technology*, edited by G. Timp, Springer-Verlag, New York, 1997.
9. B. E. Weir, P. J. Silverman, D. Monroe, K. S. Krisch, M. Alam, G. B. Alers, T. W. Sorsch, G. L. Timp, F. H. Baumann, C. T. Liu, Y. Ma, and D. Hwang, "Ultra-Thin Gate Dielectrics: They Break Down, but Do They Fail?," *IEDM Technology Digest* (forthcoming).
10. K. P. Cheung and C. S. Pai, "Charging Damage from Plasma Enhanced TEOS Deposition," *IEEE Electron Device Letters*, Vol. 16, No. 6, 1995, p. 220.
11. P. K. Roy and I. Kizilyalli, "Synthesis and Characterization of a Stacked SiO_2-Ta_2O_5-SiO_2 Gate Dielectric for Giga-Scale Integration of CMOS Technologies," *Applied Physics Letters*, (forthcoming).
12. B. J. Sapjeta, T. Boone, J. Rosamilia, P. J. Silverman, T. W. Sorsch, G. L. Timp, and B. E. Weir, "Minimization of Interfacial Microroughness for 13-60Å Ultra-Thin Gate Oxides," *Proceedings of the 1997 Spring MRS Meeting*, Vol. 477, San Francisco, California, Apr. 1-3, 1997, p. 203.
13. K. S. Krisch, J. D. Bude, and L. Manchanda, "Gate Capacitance Attenuation in MOS Devices with Thin Gate Dielectrics," *IEEE Electron Device Letters*, Vol. 17, No. 11, Nov. 1996, pp. 521–524.
14. M. L. Green, D. Brasen, K. W. Evans-Lutterodt, L. C. Feldman, K. Krisch, W. Lennard, H. T. Tang, L. Manchanda, and M.-T. Tang, "Rapid Thermal Oxidation of Silicon in N_2O Between 800 and 1200° C: Incorporated Nitrogen and Interfacial Roughness," *Applied Physics Letters*, Vol. 65, No. 7, Aug. 15, 1994, pp. 848–850.
15. K. K. Ng and W. T. Lynch, "Analysis of the Gate-Voltage-Dependent Series Resistance of MOSFETs," *IEEE Transactions on Electron Devices*, Vol. ED-33, No. 7, July 1986, pp. 965–972.
16. J. D. Bude, "Gate Current by Impact Ionization Feedback in Submicron MOSFET Technologies," *1995 Symposium on VLSI Technology*, Kyoto, Japan, June 6-8, 1995, pp. 101–102.
17. G. Timp, A. Agarwal, F. H. Baumann, T. Boone, M. Buonanno, R. Cirelli, V. Donnelly, M. Foad, D. Grant, M. Green, H. Gossmann, S. Hillenius, J. Jackson, D. Jacobson, R. Kleinman, A. Kornblit, F. Klemens, J. T.-C. Lee, W. Mansfield, S. Moccio, A. Murrell, M. O'Malley, J. Rosamilia, B. J. Sapjeta, P. Silverman, T. Sorsch, W. W. Tai, D. Tennant, and B. E. Weir, "Low Leakage, Ultra-Thin Gate Oxides for Extremely High-Performance Sub-100 nm nMOSFETs," *IEDM Technical Digest* (forthcoming).
18. K. F. Lee, R. H. Yan, D. Y. Jeon, G. M. Chin, Y. O. Kim, D. M. Tennant, B. Razavi, H. D. Lin, Y. G. Wey, E. H. Westerwick, M. D. Morris, R. W. Johnson, T. M. Liu, M. Tarsia, M. Cerullo, R. G. Swartz, and A. Ourmazd, "Room Temperature 0.1 μm CMOS Technology with 11.8 ps Gate Delay," *IEDM Technical Digest*, Dec. 1993, pp. 131–134.
19. M. R. Pinto, C. S. Rafferty, R. K. Smith, and J. D. Bude, "ULSI Technology Development by Predictive Simulations," *Proceedings of the IEEE International Electron Devices Meeting*, Washington, D.C., Dec. 5-8, 1993, pp. 701–704.
20. J. B. Burr and J. Shott, "200mV Self-Testing Encoder/Decoder Using Stanford Ultra-Low-Power CMOS," *Proceedings of the 1994 IEEE International Solid-State Circuits Conference*, San Francisco, California, pp. 84–85.
21. F. Assaderaghi, D. Sinitsky, S. Parke, J. Bokor, P. K. Ko, and C. Hu, "Dynamic Threshold Voltage MOSFET (DTMOS) for Ultra-Low Voltage Operation," *Proceedings of the 1994 IEEE International Electron Devices Meeting (IEDM)*, San Francisco, California, pp. 809–812.
22. R.-H. Yan, A. Ourmazd, K. F. Lee, D. Y. Jeon, C. S. Rafferty, M. R. Pinto, "Scaling the Si Metal-Oxide-Semiconductor Field-Effect Transistor into the 0.1-μm Regime Using Vertical Doping Engineering," *Applied Physics Letters*, Vol. 59, No. 25, Dec. 16, 1991, pp. 3315–3317.
23. I. Y. Yang, C. Vieri, A. Chandrakasan, and D. A. Antoniadis, "Back Gated CMOS on SOIAS

for Dynamic Threshold Voltage Control," *IEDM Technical Digest*, Dec. 1995, pp. 877–880.
24. J. D. Bude, A. Frommer, M. R. Pinto, and G. Weber, "EEPROM/flash Sub-3.0V Drain-Source Bias Hot Carrier Writing," *IEDM Technical Digest*, 1995, pp. 989–991.
25. Z. S. Gribnikov, K. Hess, and G. A. Kosinovsky, "Nonlocal and Nonlinear Transport in Semiconductors: Real-Space Transfer Effects," *Journal of Applied Physics*, Vol. 77, No. 4, Feb. 15, 1995, pp. 1337–1373.
26. S. Luryi, P. M. Mensz, M. R. Pinto, P. A. Garbinski, A. Y. Cho, and D. L. Silvco, "Charge Injection Logic," *Applied Physics Letters*, Vol. 57, No. 17, Oct. 22, 1990, pp. 1787–1789.
27. J. Morton, "From Physics to Function," *IEEE Spectrum*, Vol. 62, 1965, p. 134.
28. F. Capasso and R. A. Kiehl, "Resonant Tunneling Transistor with Quantum Well Base and High-Energy Injection: A New Negative Differential Resistance Device," *Journal of Applied Physics*, Vol. 58, No. 3, Aug. 1, 1985, pp. 1366–1368.
29. T. A. Fulton and G. J. Dolan, "Observation of Single-Electron Charging Effects in Small Tunnel Junctions," *Physical Review Letters*, Vol. 59, No. 1, July 6, 1987, pp. 109–112.
30. A. Dodabalapur, A. J. Laquindanum, H. E. Katz, and Z. Bao, "Complementary Circuits with Organic Transistors," *Applied Physics Letters*, Vol. 69, No. 27, Dec. 30, 1996, pp. 4227–4229.
31. S. Hauck, "Asynchronous Design Methodologies: An Overview," *Proceedings of the IEEE*, Vol. 83, No. 1, Jan. 1995, pp. 69–93.
32. U. Ko, A. Pua, A. Hill, and P. Srivastava, "Hybrid Dual-Threshold Design Techniques for High-Performance Processors with Low-Power Features," *Proceedings of the 1997 International Symposium on Low Power Electronics and Design*, Monterey, California, 1997, pp. 307–311.
33. A. E. Dunlop, W. J. Evans, and L. A. Rigge, "Managing Complexity in IC Design—Past, Present, and Future," *Bell labs Technical Journal*, Vol. 2, No. 4, Autumn 1997, pp. 103–125.

(Manuscript approved October 1997)

WILLIAM F. BRINKMAN is physical sciences research vice president at Bell Labs in Murray Hill, New Jersey. He received B.S. and Ph.D. degrees in physics from the University of Missouri in Columbia, and subsequently spent one year as a National Science Foundation Postdoctoral Fellow at Oxford University in England. Early in his career, Dr. Brinkman worked on theories of condensed matter, which included the theory of spin fluctuations in metals and other highly correlated Fermi liquids. This work resulted in a new approach to highly correlated liquids in terms of almost localized liquids. Later, he explained the superfluid phases of one of the isotopes of helium and many properties of the exotic states of matter. He co-developed the theoretical explanation of the existence of electron-hole liquids in semiconductors. His theoretical work on liquid crystals and incommensurate systems contributed to the theoretical understanding of condensed matter.
Dr. Brinkman is a fellow in the American Association for the Advancement of Science and the American Physical Society. He chaired the National Academy of Sciences Physics Survey and the organization's Solid-State Committee. He served on the Council of the National Academy of Sciences and is a member of the American Academy of Arts and Sciences. For his own research and his research leadership, Dr. Brinkman received the 1994 George E. Pake Prize.

MARK R. PINTO is the newly appointed chief technical officer and vice president of the Microelectronics Group of Lucent Technologies in Berkeley Heights, New Jersey. He has B.S. degrees in both electrical engineering and computer science from the Rensselaer Polytechnic Institute in Troy, New York, as well as M.S. and Ph.D. degrees in electrical engineering from Stanford University in California. As part of his doctoral work, he developed the semiconductor device simulation program called PISCES-II, a standard development tool in the IC industry for more than ten years. In his new position, Dr. Pinto directs strategic technology planning and implementation of new technology-based business opportunities. Earlier at Bell Labs, he conducted research in silicon IC technology and was recognized for his work in semiconductor device physics and computational simulation. He subsequently led efforts in CMOS IC process technology and circuit design research. Shortly thereafter, he became director of the Silicon Electronics Research Laboratory, which conducts advanced R&D in materials, processes, and devices, as well as in IC design. He has authored and coauthored more than 140 journal and professional conference papers. Dr. Pinto is a Bell Labs Fellow and a senior member of the IEEE. ◆