Preface

This issue, "Scaling CMOS to the Limit," comes at a point in the evolution of CMOS where, on the one hand, the technology has achieved the status of a prime mover of society, the "silicon age," in which past visions of a dollar per MIPS (million instructions per second) of computer power and a dollar per megabyte of memory have come to pass; while on the other hand the limits of CMOS scaling can be clearly seen. IBM, a prime contributor to this revolution, has seen its computers shrink from the size of a room to a thumbnail-sized chip. This enormous downscaling has been guided, all the while, by the scaling theory of IBM's Robert Dennard. On the other hand, the exponential growth in integrated circuit complexity, which has seen a hundred-millionfold increase in transistor count per chip over the past forty years, is finally facing its limits. Limits projected in the past have seemed to melt away before the concerted efforts of researchers and technologists, yet this time the limits seem more real and already are forcing new strategies on the design of future devices.

This issue of the *IBM Journal of Research and Development* contains papers by distinguished authors, both from within and outside IBM, which describe where we are today and present options for the road ahead. The future calls for major changes, ranging from device design to system implementation, in contrast to the more evolutionary path pursued up to now. A recent major change in device design was the introduction by IBM of silicon-on-insulator (SOI) technology (Shahidi). Even more radical changes in device structures and new materials may be in the offing (Wong, Nowak). Bipolar CMOS presents new opportunities for mixed-signal systems (Ning). The dynamic random-access memory (DRAM) (Mandelman and Dennard) has long had to cope with limits, and presents an instructive example of how exponential progress can be maintained, despite limits, through major innovations in device structure. The DRAM has also taught us the value of redundancy, lessons which will increasingly be applied to logic as well.

Even without revolutionary change, transistors at 20-nm gate length have been demonstrated and can be designed to be manufacturable (Taur). At this design point, the 50-nm node of the *International Technology Roadmap for Semiconductors* (ITRS), to be reached in ~2010, densities in billions of transistors per chip will be possible. This wealth of transistors will be used to advantage in both memory-rich and logic-rich applications. Future applications will employ multiple processing units on a chip, as well as heterogeneous systems on a chip featuring both digital and analog circuits (Nair). The large density advantage of DRAM compared to static RAM makes the use of embedded DRAM attractive. When examining the ITRS roadmap, one notices that the low-power devices have longer gate lengths. This is an illustration of application-dependent scaling limits (Frank) where leakage-tolerant, high-performance circuits can be scaled to smaller gate lengths. Scaling can also be extended by lowering the temperature, benefiting ultrahigh-performance systems.

The interconnect bottleneck is an important design challenge for future chips (Meindl), in terms of both performance and power, where new methodologies in optimizing multi-tier wiring stacks can lead to significant improvements. Exploiting the third dimension using multiple layers of devices might be the way to go once planar structures have been scaled to the limit, though cooling is an unsolved problem. IBM has developed an impressive interconnect technology, which was one of the subjects of another issue of the *Journal* (Volume 44, Number 3).

Predicting the reliability of future chips is difficult. In the field of oxide integrity, there is considerable difficulty even in defining what constitutes failure (Stathis), let alone elucidating the failure mechanisms. However, guidelines based on analysis of extensive data (Wu), and extrapolation of these data to a region in which direct experiments cannot be done, allows one to hope that future product criteria can be met. It is certain, however, that the gate leakage current at the 50-nm roadmap node (0.5–0.8 nm) cannot be met with silicon dioxide as the gate insulator.

Reaching the 50-nm node will present major challenges (Osburn). While CMOS technology has been very conservative in its rate of introduction of new materials, research today is being carried out on many fronts: new gate dielectrics to reduce gate leakage, metal gates to eliminate polysilicon depletion, and epitaxial silicon or silicon–germanium raised source/drains to improve series resistance. Strained-silicon concepts, using silicon–germanium to induce the strain, can improve device performance.

Maintaining dimensional integrity at the limits of scaling is a challenge. Defining patterns by lithography is itself a challenge, as described in Volume 45, Number 5 of the *Journal*; beyond this, however, processes will be required approaching atomic-layer precision (Agnello). Just being able to model future processes to predict geometries and doping concentrations of future devices is a challenge that has not been met. Our empirical techniques will have to be aided by increasingly sophisticated *ab initio* calculations in order to reduce the experimental parameter space to manageable proportions (Law). Device models, also, while reliable, are largely empirical. Careful measurements and interpretation of transport properties in small FETs (Lochtefeld) show that there is performance in silicon CMOS that has not yet been fully exploited.

In lieu of conventional scaling, more radical device and circuit concepts may extend the technology (Nowak, Wong). This includes the use of strained silicon. Today there is extensive research into double-gated FETs (the "ultimately scalable FETs"), which have better electrostatic integrity and theoretically have better transport properties than the single-gated FETs that have served us until now. Many innovative structures, involving structural challenges such as fabrication on nanometer-scale fins and nanometer-scale planarization over an entire wafer, are under experiment. Some new and revolutionary technology such as nanotubes or molecular transistors might be on the horizon, but it is not clear, in view of the predicted future capabilities of CMOS, that it will be competitive. Thus, while the traditional approaches to CMOS scaling are being pushed to their limits, new devices, materials, and designs are being developed to further extend the "silicon age."

P. M. Solomon
IBM Thomas J. Watson Research Center
Yorktown Heights, New York

Guest Editor