

FEATURE

As certain critical transistor parameters are scaled down, other parameters must also shrink accordingly

Semiconductors Shrink into the 21st Century

by Yoshio Nishi

In the semiconductor industry, looking ahead is a way of life. Because the pace of change is exponential, companies must know where the technology will be years from now in order to survive.

But looking ahead 25 years is quite a challenge. The Semiconductor Industry Association (SIA) has a projection, or roadmap, but it extends only to 2010, a mere 13 years from now. And when one considers that 50 years after the invention of the transistor, manufacturers are able to pack more than a million transistors on a chip smaller than the original device, one has to wonder how accurate any 25-year prediction can be.

In the past, the major driver for scaling has been the need to achieve the highest density and greatest cost reduction per logic function or memory bit. Several integrated circuit (IC) generations ago, Gordon Moore of Intel predicted that the number of memory bits on a chip would quadruple every three years. Because this prediction has proved true so far, it has come to be known as Moore's Law and is used as a yardstick for future progress in integration. If Moore's law continues to hold, 16-terabit dynamic random access memory chips (DRAMs) and ICs of equivalent complexity will be in production in 2022.

Before we fantasize about what a terabit computer could do, however, we need to consider another rule of enormous practical importance: the scaling principle, which states that as certain critical transistor parameters are scaled down, other parameters must shrink accordingly. For example, as transistors become smaller, the silicon dioxide layer that insulates the transistor's gate from the channel beneath it must become thinner. Sustaining the rate of progress predicted by Moore's Law means overcoming fundamental physical obstacles imposed by the scaling principle. For example, if the gate oxide were only atoms thick, it would be difficult to create a uniform layer on the wafer (even supposing the tunneling current through a thin oxide layer could be tolerated).

The challenges in the years ahead will include improv-

ing processing technologies, particularly lithography, solving the materials problems posed by layered conductive interconnects and insulators, devising methods of connecting devices into circuits that reduce complexity to a manageable level, and imagining new circuit architectures and systems that will make the best use of the computing power that continued scaling will allow.

But before we discuss these challenges, it is worth touching on two fundamental issues. One is whether future ICs will still be made of silicon. To date, the semiconductor industry has found that where silicon can do the job, the enormous capital investment in this material prevents other materials from making inroads on its territory. The manufacturing question during the next 25 years will be how long we can continue to pack more transistors on silicon, not whether we should be using another material.

Another fundamental question is how the vastly increased number of transistors that continued scaling will bring can best be used. So far, applications have swallowed up transistors as fast as the industry can make them, and human ingenuity will probably not fail us in the next 25 years. But if scaling continues, tomorrow's electronic systems will be as different from today's computers and communications systems as these are from crystal radios.

Making transistors

The greatest challenge in creating the transistors themselves will be to improve lithography—that is, to improve the techniques used to transfer the patterns that define circuit elements onto single-crystal silicon.

Most digital ICs consist of complementary metal-oxide-semiconductor (CMOS) logic gates made up of two types of metal-oxide-semiconductor field-effect transistors (MOSFETs). The attraction of this technology is its low power consumption: the transistors are paired in such a way that current flows only during the switching operation itself. The SIA roadmap calls for shrinking transistor geometries by a factor of 0.7 every three years, reaching 0.07- μm gate lengths in 2010, four IC generations from now. Several studies suggest that sub-0.1 μm CMOS devices will work well at room temperature. But if device physics does not pose an immediate problem, we are just beginning to face the production problems raised by transistors this small.

Dr. Yoshio Nishi will deliver a talk on the theme “Future development of the semiconductor electronics industry” at the annual AIP Corporate Associates meeting, hosted this year by Texas Instruments (TI) in Dallas. The meeting, which will be held October 27-28, will examine the future of the networked society being created by semiconductor electronics. Other theme session speakers will address such topics as network technology, multimedia signal processing, and network/human interactions.

The meeting will expose participants to digital light-pro-

cessing technology being developed by TI and will include tours of a state-of-the-art R&D facility at TI. An after-dinner speech on “Digital Connectivity in the 21st Century” will be given at a banquet in the Dallas Science Place Museum.

Meeting participants will also hear a lively discussion of intellectual property rights and the Internet, and lectures from scientists working at the cutting edge of physics in such areas as atomic lasers, nanotubes,

magneto-electronics, and quantum optics. Additional information about the upcoming meeting, including details about the preliminary program and on-line registration, is available on the Corporate Associates homepage (<http://www.aip.org/aip/corporate>).

The reports from the 1996 and 1995 meetings can also be found on the homepage. Last year, the AIP Corporate Associates meeting was hosted by Schlumberger-Doll Research (SDR) in Ridgefield, CT. The meeting explored the scientific, economic, and environmental challenges of satisfying the world's growing energy demand.

Until now, transistors have been defined by photolithography. The IC's structure is divided into horizontal layers, and each layer in turn is copied onto a wafer coated with a photosensitive polymer by shining light through a patterned mask or reticle. Because the smallest element on the chip cannot be smaller than the exposure wavelength, photolithography has migrated toward shorter and more expensive wavelengths with each succeeding IC generation.

Extended ultraviolet wavelengths might remain practical, although expensive, for several chip generations to come. Alternatively, the industry might shift to soft X-rays, which are much cheaper to produce. The problem with X-rays, however, is that the pattern on an X-ray reticle must be the same size as the pattern on the wafer—optical projection systems allow patterns on reticles to be five times larger than those on the wafer—and the industry has trouble creating perfect reticles even of today's comparatively coarse patterns. Another possibility is to write the pattern directly on the chip with an electron beam, the technique now used to make reticles. However, it takes 10 to 1,000 times longer to create a pattern with e-beam than with optical or X-ray lithography, a prohibitive penalty considering a wafer may have a dozen or more process layers. Only experimentation will tell which method of lithography will be the most practical eight IC generations from now.

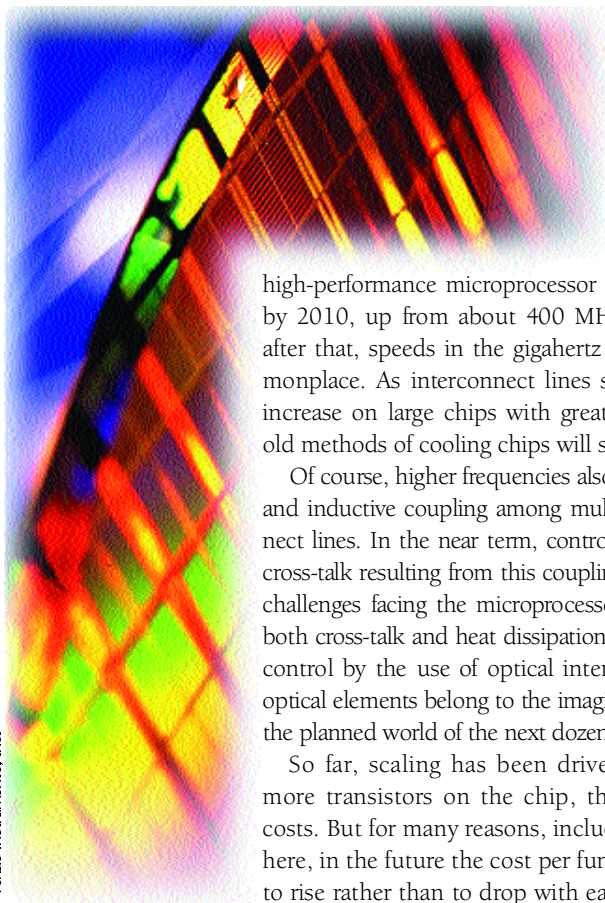
At 0.02 μm or smaller (that is, by the year 2018 or sooner) device physics will become a problem. In a MOS-FET, the imposition of a voltage on a gate electrode creates a channel beneath the gate through which current flows from a source to a drain region. When channels are only 0.02 μm long, electrons will be able to travel their length without colliding with holes or atoms, a phenomenon called ballistic transport. If electrons do not undergo collisions, their velocity does not “saturate,” instead they continue to accelerate. Since some electrons will reach saturation velocity in passing through the channel and others will not, it will be difficult to calculate the current that will flow through the channel when a voltage is applied to the gate. Indeed, at the minimum feature size the transistor's performance will become unstable. This fundamental problem may shake the dominance of FETs and the CMOS process well before 2022, forcing the industry to turn to a new physical model and new devices, such as quantum devices or true ballistic-transport devices.

Miles of interconnect

The capacity of DRAM chips will continue to increase as long as smaller transistors can be packed more densely on a chip, but when it comes to high-speed microprocessors and digital signal processors, there is the additional challenge of interconnects. Whereas today's DRAMs require only two or three layers of metal interconnect on a chip, microprocessors require five or more, and the SIA roadmap predicts that this number will rise to seven or eight by 2010. Who knows how many layers will be required to connect tera-transistor logic in a quarter century?

Leaving aside the incredible complications of routing, dense interconnects are subject to fundamental physical problems. For one thing, as aluminum interconnect lines become thinner, there will be fewer atoms in the wire's cross section. Since the lightweight aluminum atoms are easily displaced by collisions with electrons, gaps can develop in traces, creating short-circuits. Within the next few IC generations, copper will replace aluminum because its heavier atoms are more likely to stay put.

But the biggest interconnect problems are heat dissipation and cross-talk among many high-speed signals running close together. Thinner interconnect lines means higher resistivity and greater heat dissipation, as does the relentless rise in the frequency at which logic is clocked. The SIA roadmap indicates that the clock frequency of a



high-performance microprocessor will rise to 1100 MHz by 2010, up from about 400 MHz today. Twelve years after that, speeds in the gigahertz range should be commonplace. As interconnect lines shrink and frequencies increase on large chips with greater transistor densities, old methods of cooling chips will simply be inadequate.

Of course, higher frequencies also increase the capacitive and inductive coupling among multiple layers of interconnect lines. In the near term, controlling the on-chip signal cross-talk resulting from this coupling is one of the greatest challenges facing the microprocessor industry. Ultimately, both cross-talk and heat dissipation may be brought under control by the use of optical interconnects. But on-chip optical elements belong to the imagined world of 2022, not the planned world of the next dozen years.

So far, scaling has been driven by economics: the more transistors on the chip, the less each transistor costs. But for many reasons, including the ones outlined here, in the future the cost per function or bit may begin to rise rather than to drop with each new IC generation. This is all the more likely because chips may grow as large as 50 cm² in the next 25 years, and the yield of good chips is proportional to the chip's size. But even if economics brings the scaling of ICs to a halt, this will not necessarily mean that computing power will thereafter be frozen. Help may come from system-level solutions, such as parallel processing architectures, multi-level logic, and quantum computing.

Memory architectures may require radical revision as well. If today's central processing units (CPUs) suffer from megabit memory bottlenecks, how badly will the gigahertz CPUs of the future thirst for terabits? Some have proposed distributing memory among logic circuits instead of isolating it on a chip. A computer with distributed memory would more closely resemble the human brain, which doesn't have separate storage and processing areas. Although this approach might allow the construction of denser computers, they would not necessarily be faster.

Quantum leaps

A different solution to the limits of CMOS scaling is offered by quantum-effect devices. By restricting the motions of electrons in semiconductors to one wavelength in one or several dimensions, it is possible to restrict the electron's allowable energies, a degree of control that can be exploited in practical devices. Because the electron has a wavelength of 20 nm in room-temperature gallium arsenide, the favored material for quantum devices, these devices are called nanoelectronics.

Among the most promising nanoelectronic devices are resonant tunneling devices (RTDs). By sandwiching very


thin layers of different compounds, it is possible to create barriers that normally will stop the flow of electrons. Specific voltages on a control terminal will then allow electron waves to tunnel through the barriers. One potentially enormous advantage of RTDs is their multiple switching voltages. If the multiple voltages were assigned to multiple logic states, it might be possible to store or process multiple binary bits with the same device. Alternatively, the devices could be used to implement a new, higher-order logic.

One problem with RTDs is that they draw current in the off state. Power consumption is always a concern, even for CMOS, which does not draw current when off. If RTD circuits are to replace CMOS circuits, some means of bringing their power consumption down to CMOS levels must be found. And, of course, the industry will have to learn how to mass-produce RTDs economically.

Beyond the horizon

If these and other problems are solved, future computers will have capabilities quite different from today's computers. Considering their interfaces with the outside world, I believe the computers of the future will respond to people much as people respond to one another. Speech technology will be much more sophisticated than today's voice recognition systems. Keyboards and point-and-click devices will disappear—and with them the epidemic of carpal-tunnel syndrome. Computers will answer complex questions with complex answers that mirror human thinking but are perhaps more logical.

Combining sensors with ICs also offers exciting possibilities. Some interesting work is already being done to simulate in silicon the sensory functions of the nervous system, with the ultimate goal of developing integrated sensors that could serve as "prostheses" for damaged nerves. For those without nerve damage, computers may work with sensors to offer additional ways of knowing.

The myriad uses of computers 25 years from now are beyond the ability of one person to predict. They will reflect the combined ingenuity of the world's technological workforce. But all of the applications will ultimately depend on the industry's ability to meet the many foreseeable challenges of IC manufacture in the years ahead. 

Yoshio Nishi is senior vice president and director of research and development for the Semiconductor Group at Texas Instruments, Inc., Dallas, Texas.