The Amazing Vanishing Transistor Act

Radical changes are in the offing for transistors as their dimensions shrink to a few tens of nanometers

BY LINDA GEPPERT Senior Technical Editor

A decade from now you won't recognize a transistor even if it's walking toward you up the street, assuming you could see it, of course. The gate length—the marker for gauging how small that CMOS transistor is—will be roughly one-fifth the size of the smallest in production today, only 10 nm instead of today's 50 nm. To get to that size and ensure that the transistor still operates will require many changes:

- To improve performance, silicon will be mixed with a semiconductor like germanium to produce a more spacious, strained crystalline structure that lets electric charge carriers move faster.
- To reduce the leakage of current that drives up power consumption, gate oxides will be made of materials with more than eight times the dielectric constant (k) of today's silicon dioxide.
- For better control of the transistor's on and off states, gates will be of metal, instead of polysilicon.
- For better control and (again) to reduce power consumption, gates themselves will be doubled up so that two will do the job a single gate does now.

Among these techniques, strained silicon is the only one to have been commercialized so far. The rest are still at various stages of R&D. High-k dielectrics and metal gates could be next on the market as soon as they can be integrated into the manufacturing process. As for the double-gate devices, the jury is still out. Most researchers believe that they will be necessary when gate lengths shrink to 10 nm. But some think that they could be used earlier in portable applications, such as cellphones and handheld devices, to reduce the number of chips and power dissipation or to add capabilities.

Racing to the limits

Although some pundits have predicted that the evolution of semiconductor technology to smaller dimensions will slow down as dimensions shrink, things are in fact speeding up. The International Technology Roadmap for Semiconductors, published periodically by the Semiconductor Industry Association (SIA, San Jose, Calif.), recently revised its projection for the 2003 technology node from 100 nm to 90 nm. "Technology node" refers to the set of processes needed to print the smallest feature, which would be approximately 90 nm. In high-end processes, gate length may be selectively etched down to about half this minimum feature size.

True to the 2001 Roadmap projections, many foundries, including Intel, TSMC, Philips, IBM, STMicro, Motorola,

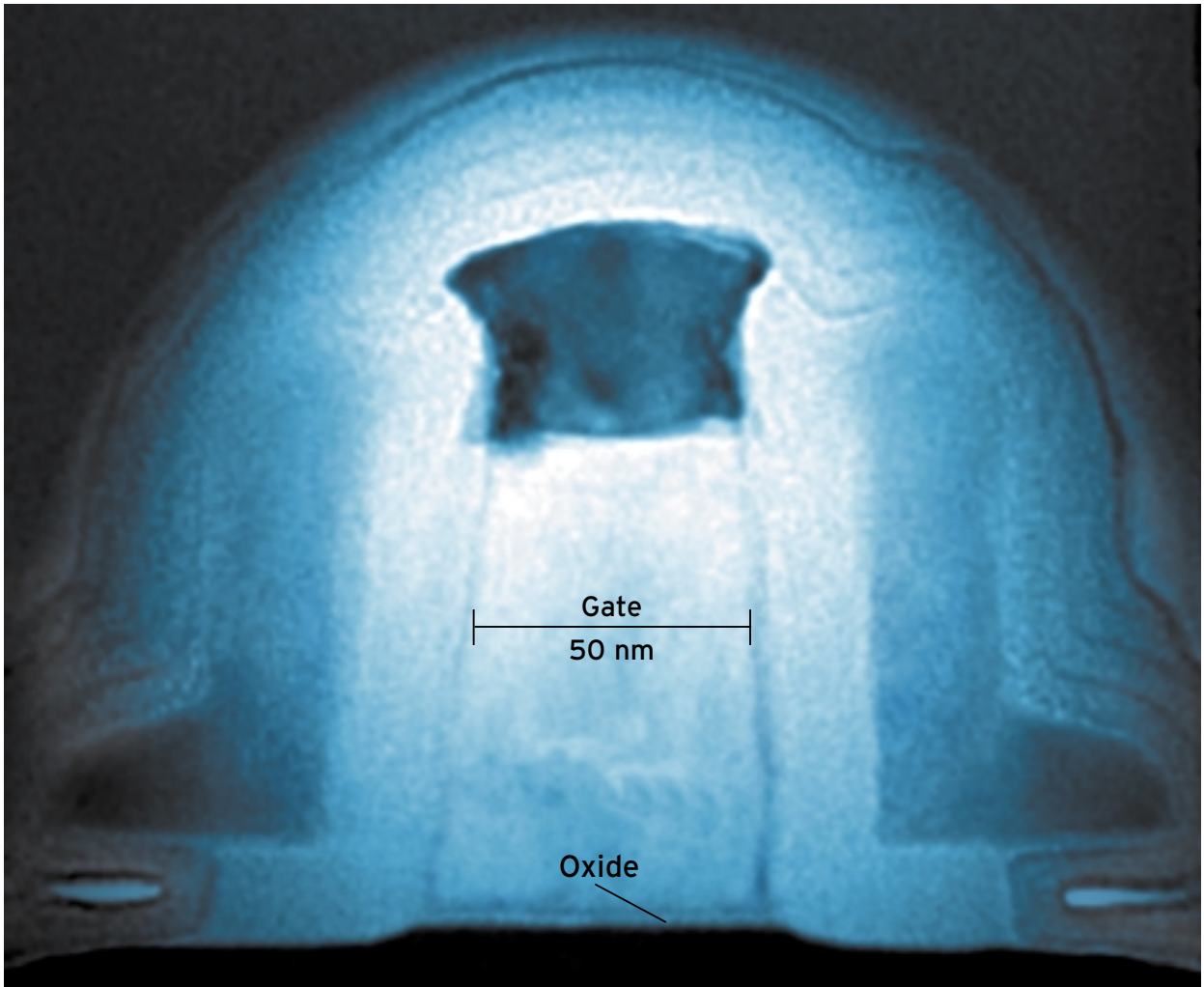
and LSI Logic, are gearing up to start volume production of 90-nm processes in 2003. Intel Corp.'s prototype 90-nm process, being brought on-line at its Beaverton, Ore., facility, has already produced a fully functional 52Mb SRAM with transistor gate lengths of 50 nm and SRAM cell sizes of just 1 μm^2, or roughly half the cell size of today's most advanced SRAMs [see figure below].

The downscaling will continue. According to the SIA's roadmap, high-performance ICs will contain by 2016 more than 8.8 billion transistors in an area 280 mm^2—more than 25 times as many as on today's chips built with 130-nm feature sizes. Typical feature sizes, which are also referred to as linewidths, will shrink to 22 nm, less than one-fifth the width found in the current generation of ICs, and power dissipation on high-performance microprocessors will double its present value, requiring more elaborate heat sinks. Some of the extra power consumption will come from gate-to-substrate and source-to-drain current leakage that will grow larger as channel lengths scale down to a few tens of nanometers.

Almost nine billion transistors on a chip may sound like a pipe dream, but scientists are already devising ways to make it a reality. Advances must be made on all fronts, including the chip manufacturing process, circuit architecture, and design methods. But no area is more essential to the future of semiconductor technology than the transistors used to build CMOS circuits [see "The Gate Rules," p. 31, for some basics of CMOS transistors].

A Catch-22

As transistor dimensions continue to shrink, the channel length (the distance between the source and drain) will shrink as well—from today's 50 nm to the 10-nm lengths expected in the next decade. A shorter channel means faster transistor switching because the charge carriers have a shorter distance to travel. But at the same time, it becomes harder for the gate to maintain control over the channel. Instead, the voltage on the drain begins to lower the energy barrier in the channel, reducing the threshold voltage and



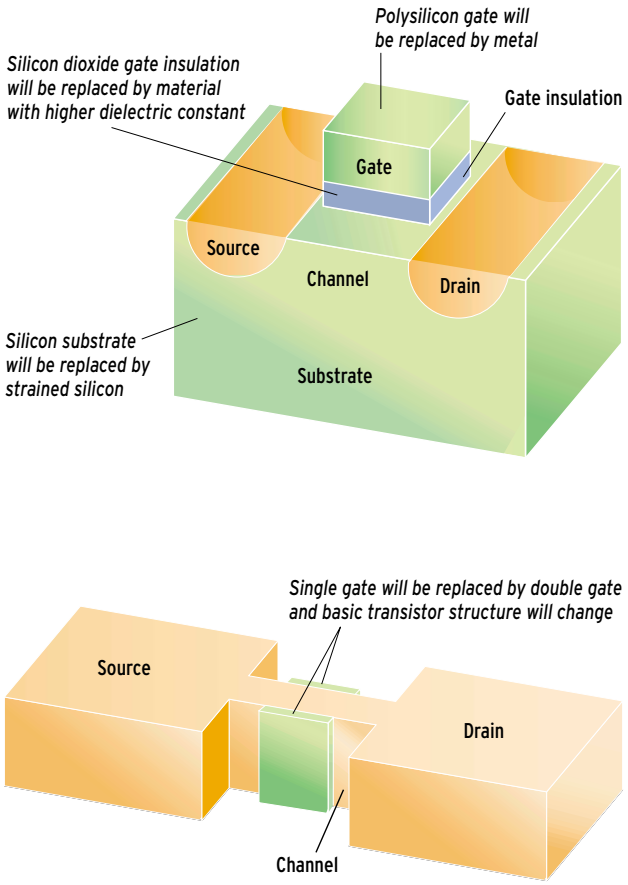
Intel is charging into the nanoscale era with its new 90-nm manufacturing process, expected to ramp up early next year. In this process, the transistors, such as the one shown in this scanning electron micrograph, will have gate lengths of only 50 nm. The silicon dioxide insulation, barely visible beneath the gate, is only about 5 atomic layers thick.

INTEL CORP.

freeing carriers to flow even when there is no voltage on the gate. This is what, in essence, is called the short-channel effect; it causes power consumption to rise and ultimately destroys transistor-switching action completely.

To keep the short-channel effect at bay, device designers must sacrifice some transistor performance and endure some increase in power consumption. Reducing the thickness of the depletion region under the gate by increasing the doping in the channel maintains gate control, but it also reduces carrier mobility (a measure of the speed with which carriers move through the semiconductor under the influence of an electric field). When the engineers also decrease the thickness of the silicon-dioxide gate insulation atop the channel to give the gate better control over the channel, the thinner oxide lets more current leak between the gate and the substrate, driving up power consumption.

Up to now, these solutions, which have also required ever more complex doping profiles, have worked well enough. But they are running out of steam. So engineers are coming up with other techniques to keep transistor performance up to par in future transistor generations [see illustration below].



The Coming Thing in Transistors

Future generations of transistors will not only be smaller than today's, but will also have to be different in more fundamental ways, as indicated above, to maintain acceptable performance levels.

Strain's the game

The technique for improving transistor performance that appears to be closest to commercialization puts the crystalline silicon on the surface of a wafer under special strain. Intel is using the technique at its Oregon fabrication facility. AmberWave Systems Corp. (Salem, N.H.) developed a version of strained silicon based on work done at Bell Laboratories and the Massachusetts Institute of Technology, and in 2001 began licensing it to wafer and chip manufacturers. It is working with semiconductor fabrication facilities to integrate the technology into their processes. IBM is also pursuing strained silicon and plans to introduce the technique in a 65-nm process, according to IBM spokesperson Michael Loughran.

Transistors built on strained-silicon wafers have shown strikingly greater charge-carrier mobility than those using conventional substrates. At this year's VLSI Technology Symposium (Honolulu, Hawaii), IBM reported increases of 60 percent. Drive current—the output current of the CMOS device—has also risen. Drive current is proportional to carrier mobility, but depends on other factors, like device geometry and capacitance, as well. A larger drive current boosts the switching speed of the transistor, so that circuits run at higher clock rates. AmberWave is seeing a 25–30 percent increase in drive currents for NMOS transistors and a 5–10 percent increase for PMOS. Intel sees a 10–20 percent increase.

In strained silicon, germanium atoms replace some of the silicon atoms near the wafer's crystalline surface; then a thin layer of silicon is grown on top of the silicon-germanium layer. Because germanium atoms are larger than silicon atoms, the distance between the atoms in the silicon-germanium layer is larger than it is in pure silicon. So when the top silicon layer is grown, its atoms line up with the silicon-germanium below, and it becomes strained—or stretched—in the two directions parallel to the plane of the wafer and compressed in the perpendicular direction [see figure, p. 32]. It is in this top strained-silicon layer that the transistors are fabricated.

In a pure silicon crystal, the distance between a silicon atom and its nearest neighbors is the same in all three directions. But in the strained-silicon layer, the atomic separations in the wafer plane are different from those in the perpendicular direction. This change in the crystal symmetry changes the energy band structure in the conduction and valence bands. AmberWave's cofounder and chief technology officer, Mayank Bulsara, told *IEEE Spectrum* that the effect of this change is to reduce electron and hole collisions with phonons (vibrations of the atoms in the crystal)—so-called scattering—that slow the carriers down. It also decreases the effective masses of the electrons and holes so that they are more rapidly accelerated by an electric field. (The effective mass of an electron or hole is a way of factoring in the forces exerted on the carriers by the atoms in the crystal.) Carriers with smaller effective masses have greater mobility, all else being equal.

Although IBM has elected to hold off bringing

MICHAEL R. VELLA

strained silicon into production, it may be for a good reason. When it does come on-line in the 65-nm process, according to IBM's Loughran, it will be on silicon-on-insulator (SOI) wafers. Widely used in semiconductor manufacturing for the past several years, SOI wafers have a layer of silicon dioxide insulator buried under the device layer, to reduce junction capacitance and so speed up transistor switching.

Using SOI with strained silicon enhances device performance just as it does with ordinary bulk silicon devices, according to Ken Rim, a research staff member working on strained silicon at IBM's Thomas J. Watson Research Center (Yorktown Heights, N.Y.). "In the strained-silicon device," he says, "most of the pn junction [at the substrate-drain or -source interface] is formed in the silicon-germanium layer, which has a smaller energy gap and a higher dielectric constant." The smaller energy gap means that carriers are more likely to leak uselessly into the substrate, driving up power consumption. The larger capacitance slows the transistor down. "The major advantage is that the SOI gets rid of a lot of junction area to reduce junction capacitance," says Rim. And leakage is blocked by the buried layer of silicon dioxide.

Stop that leak

Rim and his colleagues are working to combine strained silicon with a solution to another burning issue for future CMOS devices: greater gate leakage through the very thin gate oxides needed for nanoscale transistors. Ideally, the gate controls holes or electrons in the channel strictly through capacitive coupling, being separated from the channel by an oxide insulating barrier immune (ideally) to the passage of charge carriers. But for the 90-nm node and below, the thickness of the gate oxide is shrinking to less than 2 nm. Intel's 90-nm process has a gate oxide only 1.2 nm thick—just five atomic layers! An oxide this thin allows a significant amount of current to flow from gate to channel substrate, to no good purpose. Although higher power consumption can be tolerated by high-performance devices like microprocessors, it is beginning to interfere with the functioning of low-power ICs.

The solution is straightforward, at least on paper. Replace the present gate insulation, silicon dioxide, with a material having a larger dielectric constant. A material's k value is a measure of the extent to which it concentrates electric field lines. The capacitance between two conducting plates, in this case the gate and the substrate, goes up as the dielectric constant of the insulator between the two plates goes up. So a gate over a thick high-k insulator can control the channel just as effectively as one over a thinner lower-k insulator. And the thicker the insulator, the less current is leaked between the gate and the substrate. Scientists are investigating many high-k dielectrics, and one promising candidate is hafnium dioxide, whose k of about 22 allows the gate to control the channel despite the oxide being several times thicker than silicon dioxide.

As yet, though, hafnium dioxide is not quite ready for prime time, not least because it appears to degrade the mobility of the carriers in the channel. Scientists think that electrons may enter the hafnium oxide both when the transistors are being built and during their operation.

The Gate Rules

A CMOS circuit is built out of two types of transistors, NMOS and PMOS. Each contains four electrodes—source, gate, drain, and substrate—which manipulate electric charge carriers. In a single CMOS device, an NMOS and a PMOS transistor are connected in series to the power supply and are controlled by the same gate voltage. The advantage is that, ideally, the CMOS device consumes power only when it switches. NMOS and PMOS are so-called MOSFETs, or metal-oxide-semiconductor field-effect transistors.

The basic material of CMOS circuits is silicon, but it is doped with other materials, like arsenic or boron , to alter how it conducts electricity. Arsenic is an n-type impurity that releases one of its electrons into the conduction band; so in n-type silicon, electrons carry the current. Boron is a p-type impurity that contains electron vacancies, or holes, which behave like positive electric charge carriers.

NMOS transistors are built on p-type substrates and the drain and source are doped with n-type impurities. Conversely, PMOS transistors are built on n-type substrates and the source and drain are doped with p-type impurities. Between the drain and source is the channel, a thin layer of silicon under the gate that is depleted of carriers. Between the gate and the channel is a layer of insulation. The voltage on the gate turns the transistor on or off by controlling the flow of electrons or holes from the drain to the source.

In normal operation, the goal is to control the flow of charge carriers through the channel. In the case of an NMOS transistor, the drain is biased with a positive voltage, but even so an energy barrier, or voltage, in the channel keeps electrons from flowing from the source to the drain. Biasing the gate with a positive voltage gets things moving, lowering the barrier and attracting electrons into the depleted channel region, until above what is termed the threshold voltage, the carriers begin to flow in appreciable numbers through the channel from the source to the drain. When the gate voltage is brought below the threshold voltage, the flow is shut off. (PMOS devices work the same way, with appropriate changes in voltages and carrier type.) —*L.G.*

At IBM, researcher Rim and his co-workers have combined hafnium dioxide gate insulation with a strained-silicon substrate, with striking preliminary results. In transistors with hafnium dioxide insulator over strained silicon, mobility is 60 percent higher than when the hafnium dioxide overlies a conventional substrate, and 30 percent higher than when a silicon dioxide insulator overlies a conventional substrate.

Impressive as these figures are, the most important finding may be that hafnium dioxide and strained silicon are compatible. As Rim says, a high-k dielectric reduces gate leakage, while strained silicon increases the performance of a transistor by material innovation, not by downscaling. "It would be great if those two advantages could be combined without interfering," he told *Spectrum*. "Our preliminary experiment indicates we can

have high mobility and low gate leakage at the same time. So it's the best possible combination."

The need to maintain strong coupling between the gate and the channel as transistor dimensions shrink is indirectly the motivation for yet another materials change: metal gates. Today's transistors have polysilicon gates so highly doped as to be almost as conductive as metal. But when they are biased, a depletion region about half a nanometer thick forms at the surface of the gate in contact with the insulator, adding to the effective thickness of the gate oxide and so reducing coupling. "No matter how high you dope it, there's always a depletion region of about 4–5 angstroms [0.4–0.5 nm]. These days the oxides are very thin and so that is a lot," says Philip Wong, senior manager of nanoscale materials, processes, and devices at IBM's Watson Research Center. "But a metal has a lot of carriers, so the depletion region is almost nonexistent." So, all else being equal, a metal gate will control the channel more strongly than a polysilicon gate.

Metal gates should also be helpful in integrating the high-k dielectrics for the gate insulation into the manufacturing process. Researchers suspect that the dielectrics' performance may be degraded by the high temperatures—greater than 1000 °C—needed to move the dopant atoms into crystal positions normally occupied by silicon atoms in the polysilicon gate. "People have been having trouble integrating the high-k hafnium dioxide with polysilicon gates," says Wong. "With the metal-gate process temperatures are usually lower—below 600 °C. So to get high k, you may have to go to metal gates. But nobody knows the answer right now."

Choosing the right metal

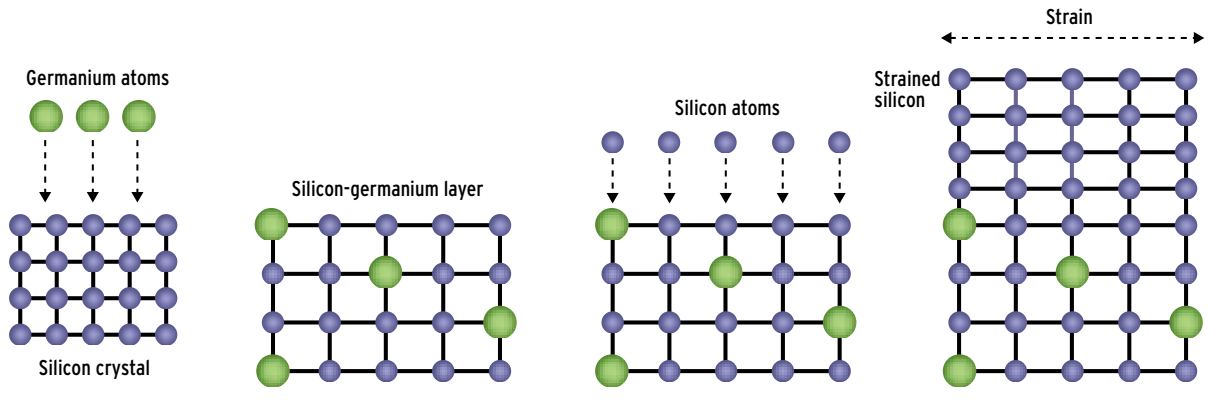
By far the biggest challenge in bringing metal gates into production is figuring out which metals to use. Several factors are at play, says Wong: the materials and their combinations, thermal stability, and reactions with the underlying insulating oxide. To give the transistor the desired threshold voltage, the metal must also have the right work function, which is the term for the energy needed to extract an electron from the metal into a vacuum. Put another way, it is a measure of how strongly electrons are attracted by a metal. The amount of voltage that you have to apply to the gate to attract electrons into the channel depends on that work function.

To complicate matters further, NMOS and PMOS transistors would require gate materials with different work functions: smaller for NMOS and larger for PMOS. Today, researchers are exploring tungsten and molybdenum, among other metals. The use of a ruthenium-tantalum alloy was proposed as a possible gate material in a paper presented at the 2001 IEEE International Electron Devices Meeting held in December (Washington, D.C.), by Huicai Zhong, then studying at North Carolina State University in Raleigh. The beauty of this approach is that the work function can be adjusted by changing the ruthenium-tantalum mix to set the required threshold voltage of the transistor.

The end of the planar road?

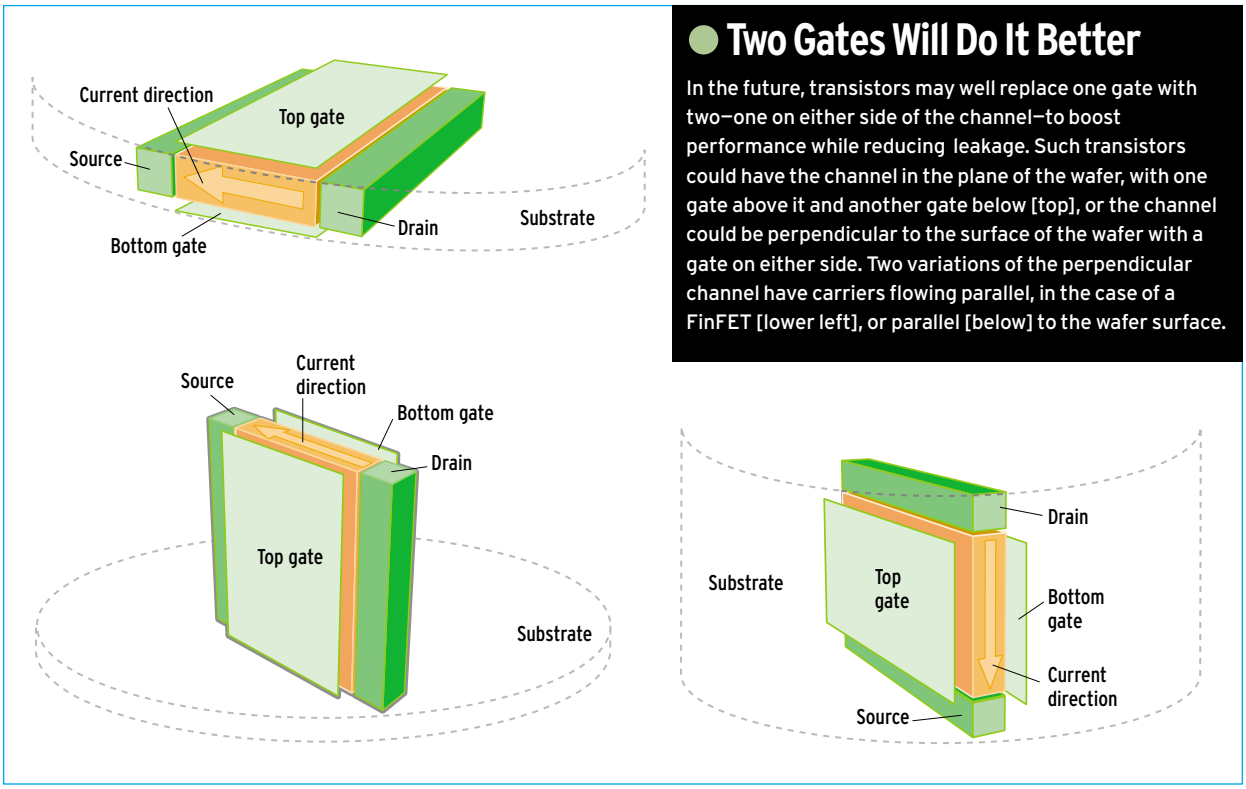
The use of high-k dielectrics, plus metal gates, plus strained silicon, plus increasingly complex doping profiles will extend the life of the planar CMOS transistor for at least another decade. But then what? Researchers are looking to double-gate transistors to take over when planar CMOS finally runs out of steam. In double-gate devices, the gate is on both sides of the channel, giving much tighter control of the transistor's on and off states.

"There are in general three ways to build double-gate devices," explained IBM's Wong. "You can do it horizontally with one gate on top and another on the bottom. You can do it vertically so that the current runs perpendicular to the silicon surface, or you can do it with the channel and gate perpendicular to the surface but with the current parallel to the surface" [see figure, next page]. The last approach is called a FinFET and is the front-runner in the double-gate device race. The approach is under development at IBM, AMD, Intel, and Hitachi, along with others in the semiconductor industry, according to Jeffrey



Putting a Strain in Silicon to Speed Up Transistors

The first step in making strained silicon is to replace some of the atoms in the top layer of the silicon wafer with germanium atoms [left]. Because germanium atoms are bigger than silicon atoms, the distance between atoms in the silicon-germanium layer increases [second left]. Next, a layer of silicon is grown on top of the silicon-germanium [second from right]. The crystal structure In this top layer of silicon[far right] is strained as it stretches to line up with the silicon-germanium layer below.



Two Gates Will Do It Better

In the future, transistors may well replace one gate with two—one on either side of the channel—to boost performance while reducing leakage. Such transistors could have the channel in the plane of the wafer, with one gate above it and another gate below [top], or the channel could be perpendicular to the surface of the wafer with a gate on either side. Two variations of the perpendicular channel have carriers flowing parallel, in the case of a FinFET [lower left], or parallel [below] to the wafer surface.

Bokor, professor of electrical engineering and computer science at the University of California at Berkeley.

The device is built by thinning the silicon layer of an SOI wafer down to a few tens of nanometers, then etching it to form a narrow vertical fin that sticks up from the wafer surface. The fin, which rests on the insulator, forms the channel of the device. Source and drain electrodes are built at each end of the fin and the gate drapes over both of its sides. Work on FinFETs has been going on for some years, but the impending end of the road for planar CMOS has researchers redoubling their efforts to perfect the device. At last December's International Electron Devices Meeting, IBM presented results on a FinFET that performed every bit as well as a conventional transistor.

One advantage of the FinFET, according to Berkeley's Bokor, is that the channel is undoped. That feature will become increasingly important as the channel length shrinks. The number of dopants in doped channels becomes exceedingly small as their length shrinks to only a few tens of nanometers. Consequently, fluctuations in this number during manufacture, along with small variations in the channel length across the chip, could wreak havoc on threshold voltages, degrading—if not ruining—circuit operation. In contrast, the absence of channel doping allows the gate to have much more influence over the device's threshold voltage.

Another advantage is that the fin can be made extremely thin. This feature means that no region of the fin escapes the influence of the gate. Power consumption is lower because there is no leakage path for charge carriers to flow along between the source and drain when the device is off.

Multiple-gate transistors may not stop at two. As we go to press, Intel researchers are preparing to describe a triple-gate transistor, also based on a silicon fin, at the International Solid State Devices and Materials Conference to be held 17–19 September (Nagoya, Japan).

TO PROBE FURTHER, SEE PAGE 67

In with the new

As for double-gate devices, Bin Yu, senior researcher at Advanced Micro Devices Inc. (Sunnyvale, Calif.), says their use will depend on the IC. "AMD is in the high-performance microprocessor business and will try to push planar CMOS to the limit," he told *Spectrum*. "But those making low-power chips—like Motorola, for example—may well use double-gate devices first, because of their incomparable ability to control current leakages, which is important for handheld products."

Other factors in addition to short channel effects and device leakages must be dealt with as well, Yu continues. With narrower linewidths and smaller source and drain junctions, the resistance in series with the transistor's channel will increase, driving up power consumption and degrading performance. With many more transistors running at much higher frequencies, IC power consumption will also rise. The amount of heat that transistors can tolerate during the manufacturing process will go down, making it more difficult to dope them effectively. Then there's manufacturability. "Can we still manufacture so many transistors on a chip with good enough uniformity of electrical performance?" Yu asks.

Engineers will likely solve these problems and others as semiconductor technology progresses through the technology nodes. And when CMOS transistors, planar or otherwise, can no longer be scaled down, more exotic devices like nanotubes, single-electron transistors, superconducting transistors, and molecular transistors will be vying to take their place. ●

MICHAEL R. VELLA