SOI to Rescue Moore's Law

Out of the Blue Come CMOS Silicon-on-Insulator Microprocessors

by Keith Diefendorff

Following last year's copper innovation, IBM is once again pushing the semiconductor industry into new territory by announcing that it will go into production with the industry's first SOI (silicon on insulator) based microprocessors in the first half of 1999. In combination with the company's copper interconnect technology (see MPR 8/4/97, p. 14), SOI should give IBM the fastest production process on the planet.

The main attraction of SOI is that it reduces transistor capacitance, thereby increasing speed and reducing power—significantly. IBM's SOI process provides as much as a 35% speedup over conventional bulk-silicon processes at the same lithography. Alternatively, SOI can reduce power 50% to 65% at the same clock speed. These improvements come at a bargain price, adding less than 10% to the manufacturing cost of a microprocessor.

Building CMOS devices on an insulating substrate is not a new idea. But early SOI circuits, built on quartz or sapphire substrates (SOS), failed to achieve mainstream status, due to cost and yield problems. Since then, research has focused on a buried layer of silicon dioxide (SiO_2) in a conventional silicon wafer as the best hope for an inexpensive insulating substrate. But it too has been plagued by manufacturing problems. Making matters worse, SOI circuits, despite their advantages, have proved difficult for designers to tame. These problems have conspired to keep SOI out of the mainstream. In fact, at times the problems have seemed so intractable that some manufactures have given up hope.

But IBM, among others, persevered. Because SOI's ability to reduce capacitance was so instinctively appealing, ten years ago IBM put a crackerjack research team to work knocking down SOI's problems. Now ready to claim victory, IBM is preparing to put the technology into volume production. While other manufacturers have produced small numbers of specialty SOI devices (only tens of thousands of SOI wafers are consumed each year), IBM will be the first to deploy it in high volume, putting the company at least one to two years ahead of its closest competitors. In its quest to bring SOI technology to the market, IBM has filed over 50 patents that could make it difficult for companies without a cross-license to follow suit.

Capacitance the Nemesis of Microprocessors

For microprocessor designers, capacitance is the enemy. The analog of mass in a mechanical system, capacitance resists voltage changes in electrical circuits. Because microprocessors process information by switching voltage—rapidly!—capacitance impedes processing. In addition, the energy required to charge capacitors is the primary source of power dissipation in CMOS circuits.

Of primary concern to circuit designers are the parasitic capacitances (C) inherent in MOS transistors. These capacitances, along with wiring capacitance, represent the AC load that must be charged before a gate's voltage will rise above its threshold (V_t), causing it to switch. Charging these capacitances takes time and is the primary source of signal-propagation delay through a transistor.

The intrinsic delay of a gate-loaded MOS transistor is given by $C \times V_{dd} / I_{DSat}$, where C comprises several parasitic capacitance terms and I_{DSat} is the transistor drive current in saturation. A large component of C is the junction-area capacitance (C_j), which is the capacitance between the source and drain diffusion regions and the substrate, as Figure 1 shows. Because the SOI substrate is an insulator, this term (with the exception of a small amount of sidewall capacitance) is virtually eliminated, causing the transistor to speed up—typically by 20–25% or more. Figure 2 shows the reduced capacitance of an SOI circuit, while Figure 3 shows that a gate-loaded SOI ring-oscillator is 25% faster than its bulk CMOS counterpart.

Reducing junction capacitance not only speeds the transistors, it reduces the dynamic power spent charging the junction capacitors. Dynamic power—by far the largest component of the power consumed by CMOS circuits—is equal to CV^2f. But SOI power savings can be even higher than is apparent from this equation: if you give back the SOI speedup by lowering the supply voltage (thereby reducing I_{DSat}), the dynamic power is reduced by the

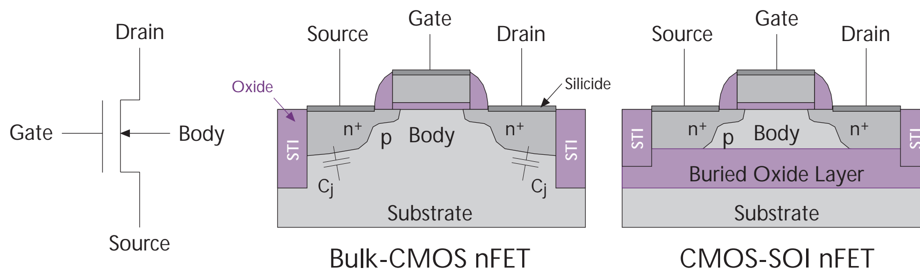

Figure 1. The structure of the bulk-CMOS transistor is identical to that of an SOI transistor, with the exception of the buried layer of silicon dioxide that insulates the SOI transistor from the substrate. The oxide layer eliminates most of the junction capacitance (C_j) but causes the body of the SOI transistor to float, introducing a number of complex circuit problems. (STI stands for shallow-trench isolation.)

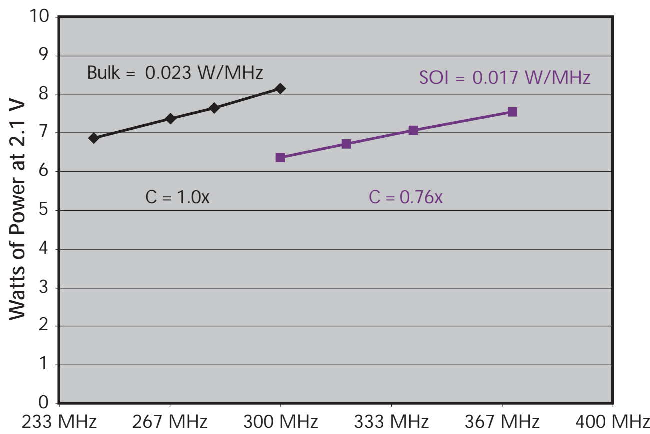

Figure 2. This graph shows the power consumption of a PowerPC 604e in 0.25-micron bulk and SOI technologies. From this data and CV^2f, the total capacitance is shown to be 24% lower on the SOI die. (Source: IBM)

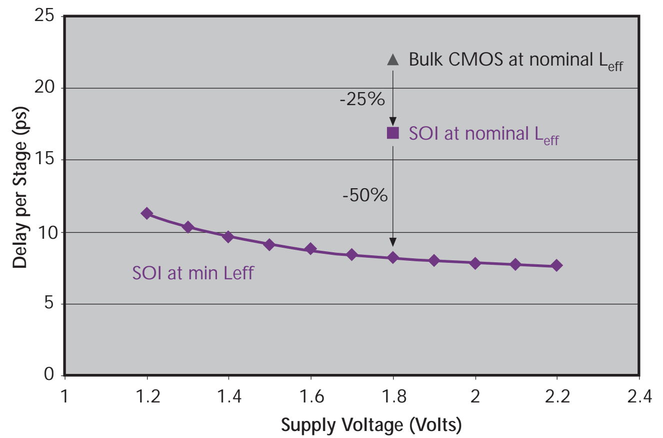

Figure 3. Unloaded ring oscillator data shows 7S-SOI transistors to be 25% faster than bulk versions at nominal L_{eff}. Pushing L_{eff} to the process minimum improves speed another 50%. Notice how little SOI's performance varies as voltage is reduced. (Source: IBM)

square of the voltage reduction. Thus a 30% reduction in capacitance, for example, can be roughly translated to a 50% power reduction at the same performance. Figure 4 shows an IBM 4-Mbit SOI SRAM delivering equivalent performance to a bulk device at one-half to one-third the power.

SOI Materials Problem Solved

The potential advantages of SOI have long been recognized. What has held back the technology is a laundry list of nasty manufacturing and circuit problems.

One of the most perplexing problems has been the manufacture of high-quality SOI wafers at reasonable cost. The difficulty is in creating a uniform-thickness, low-defect silicon layer on an insulating substrate comparable to the near-perfect surface of a single-crystal silicon wafer. Numerous schemes have failed to yield good wafers, save one: SIMOX (separation by implantation of oxygen).

In the SIMOX process, conventional silicon wafers are bombarded with a high dose of oxygen ions driven deep beneath the surface via a high-energy implant. Several hours of high-temperature annealing are applied to form the buried silicon-dioxide layer and restore the silicon layer, damaged by the implant, to its original single-crystal state. Even though the industry settled on the SIMOX process years ago, it has been far from perfect. Too far, in fact, to produce good yields on microprocessor-size die.

IBM claims to have licked this problem and can now produce SIMOX wafers of the same quality as bulk wafers. While other companies based their SOI work on wafers from external vendors, IBM developed its own secret recipe for fabricating SIMOX wafers, a step that has apparently paid off. The company is not, however, interested in the SOI wafer-supply business. IBM will depend on its own wafers for early production only because it cannot purchase wafers of sufficient quality elsewhere; it will probably turn to external vendors once they can meet IBM's standards.

According to IBM, SIMOX processing adds about 10% to the fully processed bulk-wafer cost, due mostly to the additional wafer-processing time. Implanting the 350-nm-thick oxygen layer is currently performed at about 20 wafers per day per implanter (an implanter costs about $4 million). The high-temperature anneal takes several more hours, although IBM would not disclose how many. In development are lower-dose implants that may improve implanter throughput and reduce the SIMOX cost adder to 3–5% over a conventional epitaxial wafer. As manufacturing experience is gained, better control will allow a thinner oxide layer, saving more time and further reducing the cost penalty of SOI.

Floating Body Steers Industry Off Course

SOI has historically suffered from a variety of problems arising from the floating-body effect. In bulk-silicon MOSFETs, the body of the transistor is held at a constant potential, usually ground, through contact with the substrate. As

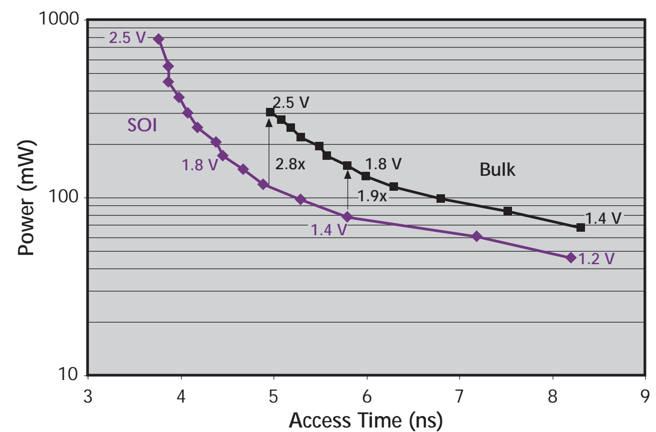

Figure 4. IBM's 4-Mbit SRAM data shows SOI's power consumption to be almost 2–3× lower than bulk CMOS's at the same speed. (Source: IBM)

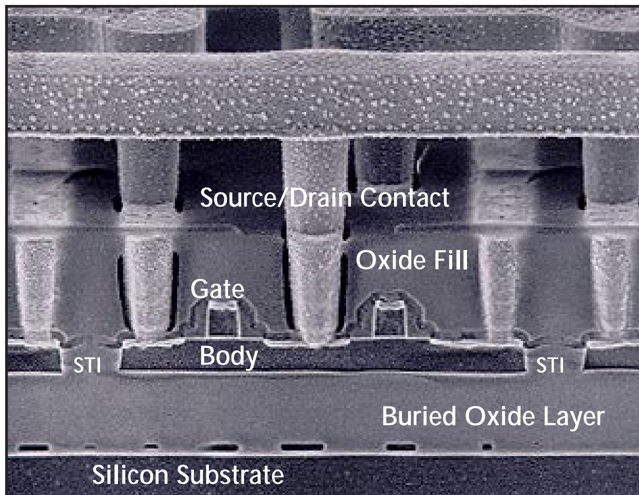

Figure 5. SOI transistors are fabricated in islands, isolated on all sides by silicon dioxide. (Source: IBM)

Figures 1 and 5 show, in SOI the body is isolated, so its potential can change. Since the threshold voltage (V_t) of a MOSFET is dependent on the body potential, a floating body can cause strange and unwanted effects.

Additional floating-body problems arise from the lateral parasitic bipolar transistor that parallels the FET. This transistor can, in the absence of a conductive substrate, get enough base drive from impact ionization current (body current arising from hot electrons impacting the drain region) to turn on, causing more problems.

To minimize floating-body effects, much SOI research has focused on fully-depleted MOSFETs. These devices have no uncharged region, giving the gate good control over the device, at least at long channel lengths. They also have a steep subthreshold-current slope, leading some to believe the V_t could be lowered (compared with bulk MOSFETs having he same leakage) to speed them up. But the better subthreshold characteristics and lower V_t advantages of fully-depleted devices disappear, and in fact reverse, at short channel lengths. Fully-depleted devices also have poor threshold control due to their sensitivity to variations in the silicon layer thickness. These problems severely limit the scalability of fully-depleted devices.

To improve scalability, IBM decided to go with partially-depleted MOSFETs. In partially-depleted devices, however, the floating-body effect precipitates a large drop in V_t at high drain voltages. To compensate, these devices must be constructed with a high V_t to reduce leakage. But a high V_t decreases performance, negating much of the advantage of SOI. This conundrum led some some researchers—notably Intel's—to conclude that below 0.25-micron geometries partially-depleted SOI MOSFETs offer little or no power or speed advantage over bulk devices.

IBM disagrees. It found that as V_{dd} is reduced, impact ionization current is reduced, thus eliminating the primary source of the V_t drop and relaxing the requirement for a higher V_t design point. Furthermore, IBM showed that while device leakage may be much higher at nominal channel lengths (L_{eff}) and room temperature (25° C), it can be matched to that of bulk devices at the more relevant conditions of minimum L_{eff} and typical junction operating temperature (85° C). Then, at the low voltages important to low-power applications (1.0–1.2 V), where the floating-body effects are less significant, the leakage will actually be less than that of bulk devices at all channel lengths, giving low stand-by power. Using its techniques, IBM sees no problem scaling partially-depleted SOI devices into the 0.13-micron realm while retaining their advantages over bulk CMOS devices.

Living With a Floating Body

IBM has overcome the problems associated with a floating body by using a combination of process adjustments and improved circuit-design tools and techniques.

One of the floating-body effects designers consider most objectionable is the kink effect—named for a sudden jump in the drain current as the drain-to-source voltage (V_{ds}) goes above a volt or so. IBM found it could live with the kink effect by accounting for it in its SOI device models and design tools. In the end, the kink effect actually turned out to be good, because it increases drive current.

Another undesirable floating-body effect is a low device-breakdown voltage. At an unusually low V_{ds}, around 2.5–3.0 volts, the SOI device breaks down, and the drain current skyrockets. This is not a problem in normal low-voltage circuit operation, but it is a nuisance in I/O driver circuits. It also hampers burn-in testing, usually performed at elevated voltages to shorten test time. IBM solved this problem simply by using lower burn-in supply voltages and extending the burn-in time to compensate, but this generally increases test costs.

The floating body also introduces an annoying history dependence into the device characteristics. Since the transistor body floats, it can accumulate a charge based on the input waveform. The resulting variation in body potential modulates the threshold voltage, causing the device speed to drift. IBM dealt with this problem by characterizing it accurately and including it, along with other sources of variation, in the worst-case delay analysis. IBM found that the history effect introduces, at most, an additional 8% delay variation above other sources of variation—such as poly-line width, temperature, and V_{dd}—that can together amount to variations on the order of 50% or more. The additional circuit-design conservatism necessary to account for the 8% history effect reduces the overall SOI speedup.

Another undesirable floating-body effect is pass-gate leakage. When the body of an SOI pass gate is charged to V_{dd} and then discharged rapidly, it sends a pulse of current through the device. Without careful circuit design, this current can upset certain types of dynamic nodes. IBM reduced this leakage current to tolerable levels with process adjustments that minimize the gain of the lateral parasitic bipolar transistor. In IBM's 0.22-micron SOI process at 2.5 V, the peak leakage current is less than 5 µA per micron of gate width.

Feature	CMOS 7S Bulk	CMOS 7S SOI	CMOS 8S Bulk	CMOS 8S SOI
Generation	0.22 µm	0.22 µm	0.18 µm	0.18 µm
Power Supply	1.8 V	1.8 V	1.5 V	1.5 V
Silicon Film Thickness	–	180 nm	–	<180 nm
Buried Oxide Thickness	–	350 nm	–	<350 nm
Metallurgy	Copper	Copper	Copper	Copper
Wiring Levels	6	6	7	7
Power (µW/MHz/gate)	0.039	0.030	0.027	0.021
Ring Oscillator Stage	22 ps	17 ps	19 ps	15 ps

Table 1. CMOS 7S-SOI is similar in most ways to its bulk counterpart but has higher speed and lower power. (Source: IBM)

While not a floating-body effect, a problem that many companies have long been worried about with SOI is self-heating. Because the silicon-dioxide layer is not a good heat conductor, it was feared that heat could build up in the transistor area, reducing its gain and robbing it of drive current. This effect, however, turns out not to be a problem. Since a CMOS device burns power only while switching, there is plenty of time for the heat to dissipate away from the device. IBM determined that self-heating raises junction temperatures by only 2–5° C worst case. Since the performance loss is only about 1.5% per 10° C, the effect of self-heating is insignificant.

Remember that as voltage is scaled down, every one of SOI's floating-body problems diminish, further increasing its advantages over bulk CMOS. For example, pass-gate leakage is negligible once V_{dd} drops below about 1.2 V. Since future processors will use ever-lower supply voltages, SOI should become even more valuable over time.

SOI Could Influence Microarchitecture

In digital circuits, it is common to stack transistors to build complex logic gates. The reverse-body effect imposes a limit on the number of transistors that can be stacked, because as the voltage difference between the source and the body decreases, V_t increases. In bulk CMOS, as transistors are stacked, the source voltage is raised but the body is held at ground by the substrate. As more transistors are stacked, the V_t of each increases, and the gate slows down. In SOI, however, the body of the transistor is not held at ground, so more devices can be stacked without V_t degradation.

This characteristic is important to microprocessor designers. To make a CPU run fast, the number of gates in each pipeline stage must be kept to a minimum. The ability to build more-complex gates (e.g. 4-input NAND vs. 2-input NAND) allows fewer gates for the same function, or more functions to be performed in the same number of gates.

So, compared with the 20–25% speedup that a CMOS SOI inverter gets over a bulk CMOS inverter, due to its lower capacitance, a complex SOI gate gets even better performance gain over bulk. For example, a four-input SOI NAND gate is 50–55% faster than the equivalent bulk version.

One profound implication of this characteristic is that microprocessor circuits need to be designed from scratch with SOI in mind if they are to take maximum advantage of the technology. While a microprocessor that was designed for bulk CMOS can be fabricated on an SOI wafer and enjoy the 20–25% speedup from its lower capacitance, this approach significantly understates the performance potential of SOI.

Furthermore, it's possible that once SOI becomes the norm, it could begin to influence processor microarchitecture or even architecture. Since the SOI speedup is not a symmetric, across-the-board speedup, it may actually favor one approach over another. This could lead the SOI haves and SOI have-nots in different microarchitectural directions. It could also become a great equalizer, allowing companies with more-complex architectures and SOI to keep pace with companies that have simpler architectures but no SOI. For example, SOI superscalar RISC processors may be able to compete in frequency with Intel's new VLIW-like IA-64 architecture (assuming Intel decides not to use SOI).

Advantages Go Beyond Speed and Power

A serendipitous benefit of SOI is its high immunity to soft errors. Soft errors occur when a high-energy particle passes through silicon, knocking loose enough electrons to upset a sensitive dynamic storage node—like the ones found in a memory cell. The critical charge necessary to upset a node (Q_{crit}) depends on the capacitance and voltage on the node. Thus, Q_{crit} drops rapidly as devices are scaled to smaller size and lower voltage, increasing the probability of soft errors.

Lower Q_{crit} could become a problem for bulk silicon, because of the large silicon substrate available to collect charge. But in SOI, with the silicon substrate insulated from the active circuity, there is much less silicon to collect charge. IBM's studies have verified that SOI's soft-error rates will indeed be better than those of bulk CMOS at 0.18 micron and below.

SOI devices also offer some advantage in packing density. Since SOI devices are isolated from the substrate, some device spacing rules can be relaxed, and direct n^+ to p^+ connections can be made. This can improve the density of transistor-dominated structures, like SRAM cells, by about 5%. In addition, without a conducting substrate there are no vertical parasitic bipolar transistors to cause latchup, as there are in bulk CMOS. This eliminates the need for epi wafers and frequent substrate ties, further improving density.

PowerPC First to Bat

IBM will first introduce SOI in its 0.22-micron CMOS-7S process, which includes copper interconnects. A nice feature of SOI technology is that the same process can be implemented on either bulk or SOI wafers with only minor adjustments. IBM will take advantage of this characteristic and offer its 7S process in both bulk and SOI variants. Bulk processing will be employed in applications where cost is more important than speed or power, and SOI will be used where speed or power are more critical. Table 1 compares the characteristics of IBM's bulk and SOI processes.

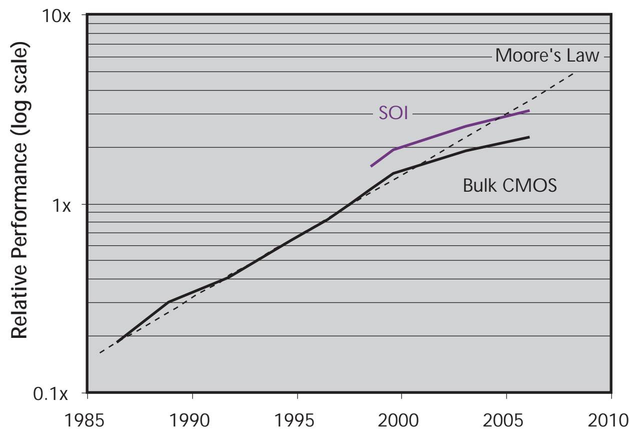

Figure 6. Bulk CMOS processes are projected to begin falling off Moore's Law curve after 0.18 micron. SOI provides a step-function improvement in speed, allowing performance to stay on track a while longer. (Source: IBM)

IBM's bulk-7S process went into production in 2Q98. The 7S-SOI process is now undergoing qualification at IBM's East Fishkill (New York) pilot line and will be moved onto the company's high-volume lines in Burlington (Vermont) in 1Q99. IBM also has plans to offer SOI versions of its 0.18-micron 8S process. The 8S process is scheduled for production in 2H99. IBM did not say when the SOI version would go into production, but we expect it to follow by about six months, putting it in 1H00. We expect that by the 0.13-micron (9S) generation, IBM will convert completely to SOI and not offer a bulk-9S process.

IBM has already demonstrated a fully functional 4-Mbit SRAM, a PowerPC 750, and a 604e in 7S-SOI. IBM would not disclose the frequency of these parts, but we expect that the SOI version of the 750 could run above 600 MHz. Some redesign to take more advantage of SOI—but still short of repipelining—might boost a 7S-SOI 750 to 700 MHz. More significant work would be required to reach 1 GHz, but SOI should help PowerPC reach this milestone.

IBM expects to enter volume production on a 7S-SOI PowerPC 750 in 1H99. This chip will be followed by a PowerPC 630FP for IBM's RS/6000 workstation line and by another PowerPC processor for its AS/400 server line. The company will also apply the technology to its System 390 mainframes and its ASICs. We expect IBM's Giga Processor to be delivered in 8S-SOI at over 1 GHz in 2000.

IBM Leads the Way

Other microprocessor vendors have not disclosed plans to deploy SOI. Motorola has said it sees benefits similar to those IBM has reported, indicating the company may be hot on IBM's heels. This is not surprising, considering the close relationship the two companies had for many years at Somerset. Both IBM and Motorola are working with Ibis Technologies (Danvers, Mass.) on high-energy oxygen implanters for SOI.

Intel, on the other hand, does not appear to be planning to deploy SOI soon, if at all. Intel says it does not see benefits from SOI like those reported by IBM; in fact Intel says that bulk CMOS is faster than SOI in its tests, especially for interconnect-dominated circuits like microprocessors. Intel is also in a different position than IBM. With that company's huge volumes, process changes can be more difficult to assimilate. For this reason, Intel rarely takes the lead on process innovations; for example, it was slow to adopt shallow-trench isolation. But if IBM's SOI claims pan out, and if AMD or other x86 competitors field SOI parts, it could put serious pressure on Intel to reconsider.

AMD declined to say anything about its SOI plans. The company has recently signed a joint-development agreement with Motorola for its 0.18-micron copper HyperMOS 6 process, but nothing about SOI was said in their announcement. If, as we believe, Motorola is readying SOI for 0.18 micron, AMD would likely receive this technology.

National also declined to disclose its SOI process plans, but it appears to be dabbling in the technology and may be interested in it, at least for its low-power attributes.

Although we have not yet seen parts, IBM's SOI claims appear to be based on solid evidence. The company's 0.22- and 0.18-micron process parameters, aside from copper and SOI, appear to be at least as aggressive as anyone else's in the industry. Adding copper interconnect and SOI technology should put IBM solidly out in front of the pack. If SOI and copper together increase performance by the 30–55% IBM claims, its 0.22-micron process could match other 0.18-micron processes in performance. This would be a truly remarkable—and significant—feat.

If IBM delivers on its SOI promises, we expect to see a multitude of other companies scramble to follow suit, as they did when IBM announced copper a year ago. In fact, after IBM's SOI announcement, we expect that many companies are already rethinking their position.

SOI Rescues Moore's Law

Semiconductor technologists have for some time been warning that the speed benefits from scaling are due to taper off as we go much below 0.18 micron. The combination of short-channel effects, difficulty in producing gate oxides much thinner than about 25 Å, and other fundamental problems may slow the speed improvements we've come to expect with each new process generation. New breakthroughs may save us from this almost unthinkable catastrophe, as they have done so many times in the past, but at least for now, slower progress in bulk-silicon speedup seems unavoidable.

Figure 6 shows IBM's projections for its next few process generations. It looks as if IBM intends to rely on SOI as the breakthrough that keeps it from falling behind Moore's Law for several more years. If others do not follow IBM down the SOI path, it will be interesting to see what they come up with to avoid being left in the dust. M