MOSFET Scaling into the Future

2D process and device simulators have been used to predict the performance of scaled MOSFETs spanning the 0.35-μm to 0.07-μm generations. Requirements for junction depth and channel doping are discussed. Constant-field scaling is assumed. MOSFET drive current remains nearly constant from one generation to the next and most of the performance improvement comes from the decreasing supply voltage. Gate delay decreases by 30% per generation, nearly the same trend as previous generations. However, this performance gain comes at the price of much higher off-state leakage because of the reduction of the threshold voltage. Various solutions to this high leakage are discussed.

by Paul Vande Voorde

Hewlett Packard adopted CMOS technology in the mid- 1970s. At that time the gate length L_g was 4 μm and the gate oxide thickness T_{ox} was 50 nm. Since then, each new generation of technology has shrunk L_g by about 30% and T_{ox} by about 25%. The decrease in L_g has been tied to the evolution of lithography equipment. Following these scaling trends, intrinsic gate delay has decreased about 30% per generation. New generations of technology are released about every three years. The important principle in MOSFET scaling is that L_g and T_{ox} must decrease together. Scaling one without the other does not yield adequate performance improvement.

The performance metric for gate delay is CV/I, where C is the load capacitance, V is the supply voltage (V_{dd}), and I is the drive current of the MOSFETs (average of NMOS and PMOS). C is composed of both gate and junction capacitance. MOSFET scaling, which decreases L_g, T_{ox}, and junction area while increasing substrate doping, tends to keep C fairly constant from generation to generation. For several generations of technology, the supply voltage was held constant at 5V (constant-voltage scaling). In that era, gate delay was reduced by ever-increasing MOSFET drive currents. Since the voltage was held constant while the dimensions decreased, the electric fields continuously increased. High fields and high currents tend to damage the gate oxide and lead to device deterioration. Thus, one of the main technology challenges has been to design MOSFETs with adequate reliability.

Constant-voltage scaling ended as L_g approached 0.5 μm and T_{ox} neared 10 nm. The demands of gate oxide reliability required that the supply voltage be reduced. This occurred as the peak oxide field reached roughly 4 MV/cm. We are now in an era where supply voltage is scaled along with T_{ox} so that the peak oxide electric field remains roughly constant (constant-field scaling). This study examines some of the implications for this of type scaling in future technology generations.

Process and Device Simulations

The 2D process simulator TSUPREM-4 from Technology Modeling Associates Inc. of Sunnyvale, California was used to simulate scaled MOSFET device structures. The inputs to TSUPREM-4 are the implant and oxidation steps that would be used in the actual process. The process architecture assumed is similar to current CMOS processes, employing shallow source/drain extensions and deeper main source/drain regions followed by silicidation.

The 2D device simulator MEDICI, also from Technology Modeling Associates Inc., was used to predict the electrical characteristics of the device structures from TSUPREM-4. Here we use field dependent mobility models that have been benchmarked to the HP CMOS10 process. Iterative simulations with TSUPREM-4 and MEDICI were performed to determine the requirements on junction depth and channel doping profile to ensure proper threshold and subthreshold behavior. Fig. 1 shows the device structures resulting from these simulations for each generation from 0.35 μm down to 0.07 μm. For L_g less than 0.15 μm, retrograde channel doping profiles are needed to control the subthreshold characteristics.

Figs. 2 through 5 summarize the results of this scaling study. Fig. 2 shows the scaling of T_{ox} with L_g. These two must scale together to get adequate performance improvement. Constant field scaling dictates that V_{dd} must decrease proportionally to T_{ox}, maintaining a peak oxide field of 4 MV/cm. For example, this results in T_{ox} = 2.5 nm and V_{dd} = 1V for the L_g = 0.1 μm generation.

Fig. 3 shows the scaling of effective channel length (L_{eff}) and the source/drain extension junction depth (X_j). For the 0.1-μm generation, L_{eff} is about 0.07 μm and X_j must be nearly 50 nm. The series resistance of the source/drain extension must

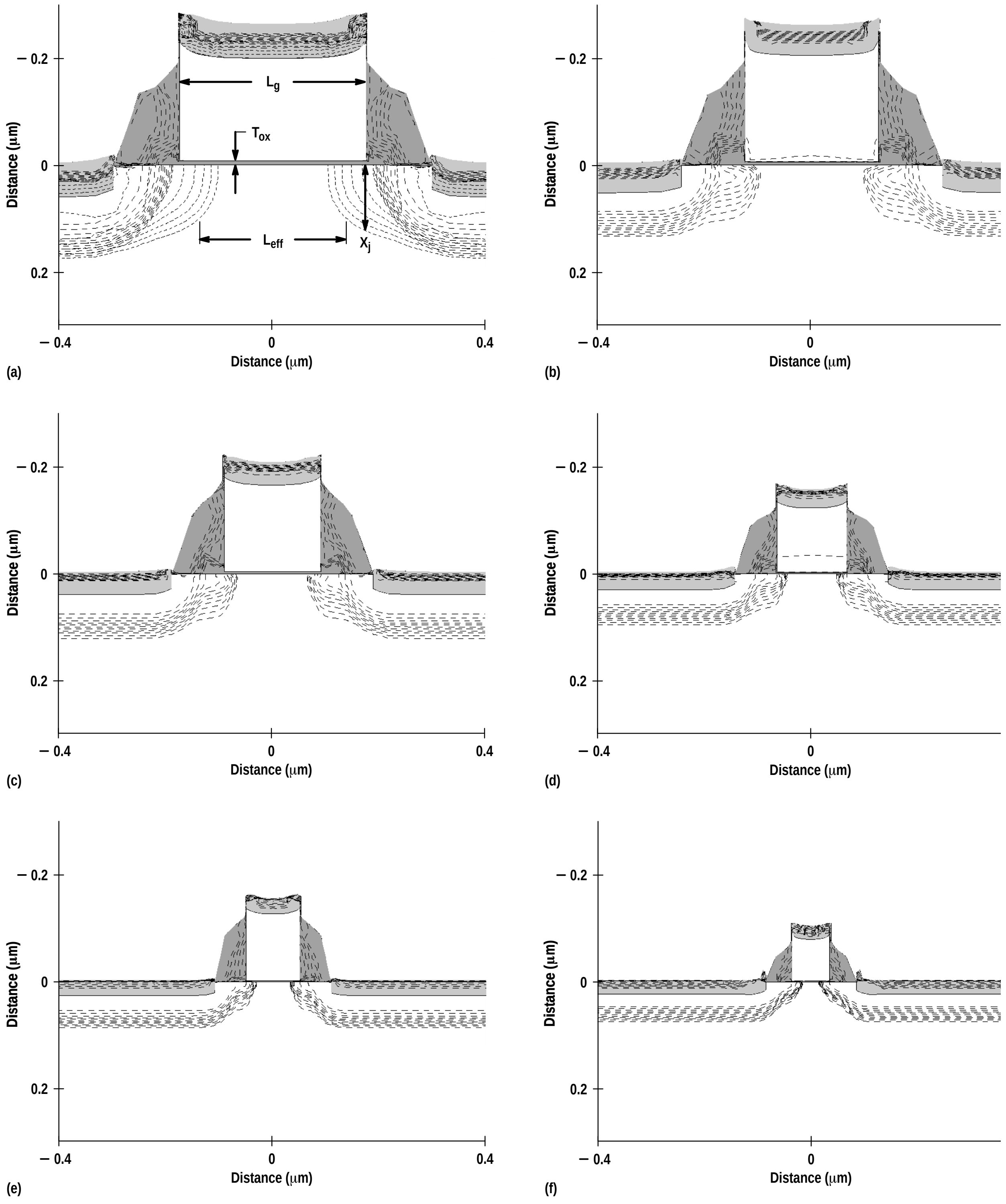

Fig. 1. *Simulated device structures. Dark shading is oxide. Lighter shading is silicide. Dashed lines are doping contours. (a) L_g = 0.35 µm, T_{ox} = 8.0 nm. (b) L_g = 0.25 µm, T_{ox} = 6.0 nm. (c) L_g = 0.18 µm, T_{ox} = 4.5 nm. (d) L_g = 0.13 µm, T_{ox} = 3.4 nm. (e) L_g = 0.10 µm, T_{ox} = 2.5 nm. (f) L_g = 0.07 µm, T_{ox} = 1.9 nm.*

decrease even as the junction depth also decreases. This requires higher doping levels in the extension region and carefully minimized spacer widths.

Fig. 4 shows the scaling of threshold voltage (V_t). Here V_t is kept at 20% of V_{dd} to maintain adequate current drive. This yields V_t = 0.2V for the 0.1-µm generation. Unfortunately, since off-state current varies exponentially with V_t, reducing V_t leads to much higher off-state leakage current (100 nA/µm for the 0.1-µm generation) than in current CMOS technologies. Here the simulations are tailored to predict the nominal leakage. Worst-case leakage would be approximately one order of magnitude higher for the 0.1-µm case.

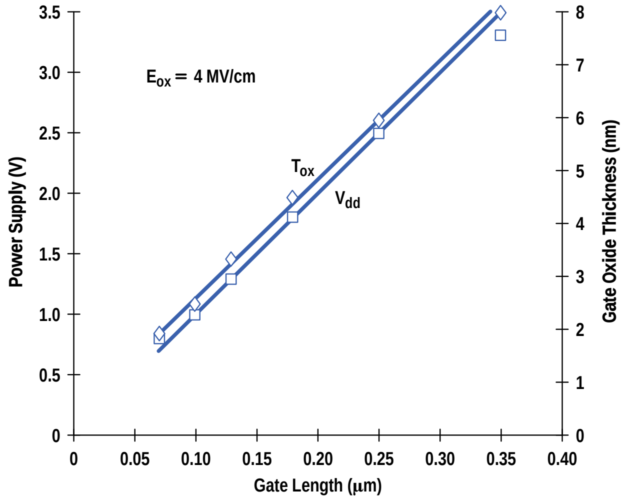

***Fig. 2.** Scaling of power supply voltage V_{dd} and oxide thickness T_{ox}.*

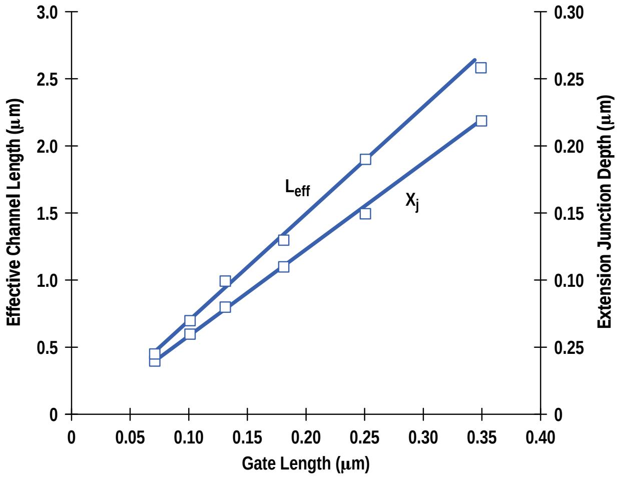

***Fig. 3.** Scaling of effective channel length L_{eff} and extension junction depth X_j.*

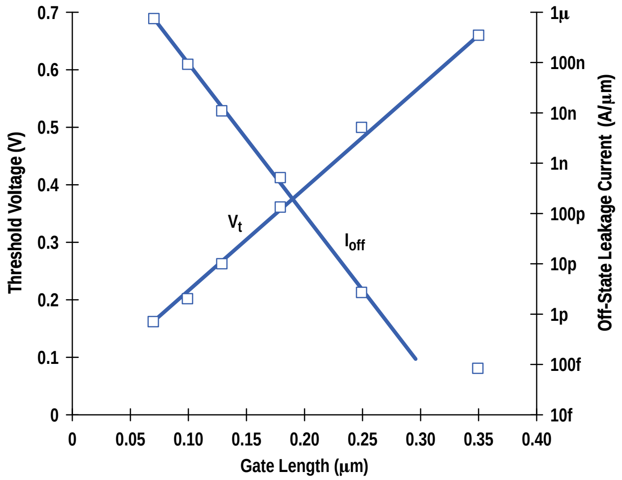

***Fig. 4.** Scaling of threshold voltage V_t and off-state leakage current I_{off}.*

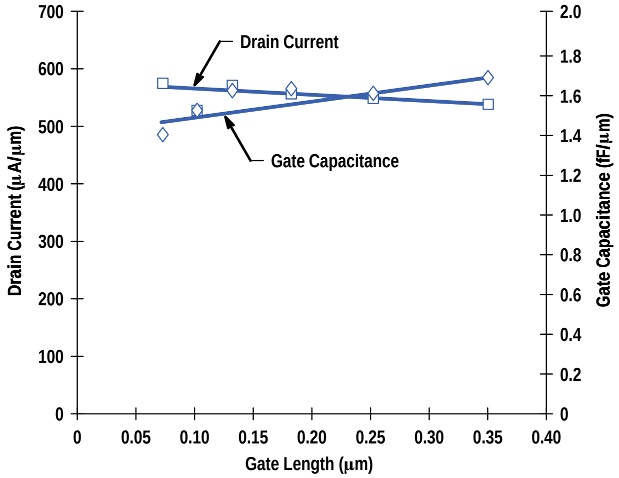

***Fig. 5.** Scaling of maximum drain current and total gate capacitance.*

Fig. 5 shows the scaling of drive current and total gate capacitance. Because of the simultaneous scaling of L_g, T_{ox}, V_{dd}, and V_t, the current and capacitance do not change much from one generation to the next. Therefore, the gate delay metric CV/I decreases primarily because of the decreasing supply voltage.

Device simulators allow one to examine the internal distributions within the device. Fig. 6 shows the lateral electric field along the channel for each of the device structures in Fig. 1. Even though V_{dd} decreases as shown in Fig. 2, the peak electric field near the drain continues to increase as L_g decreases. However, the width of the high-field region decreases, giving the electrons less and less distance to reach equilibrium with the electric field. When this “nonlocal” effect is included in MEDICI, the electron temperature can be calculated as shown in Fig. 7. Here, even though the peak field increases, the electron temperature decreases as L_g decreases. Thus, we expect that the reliability issues related to high-energy charge carriers will become less important in future generations of technology.

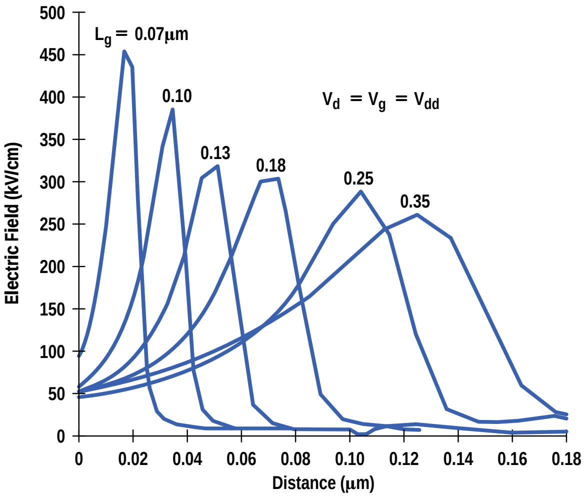

Fig. 6. *Lateral electric field along the channel beginning at the middle of the gate.*

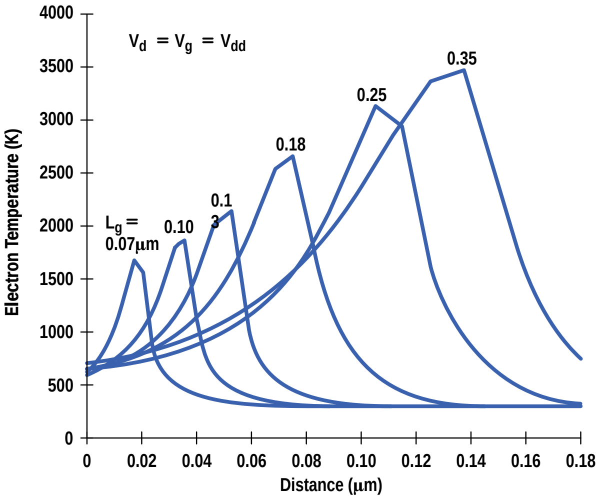

Fig. 7. *Electron temperature along the channel beginning at the middle of the gate.*

Gate Delay Simulations

MEDICI was used to generate a full set of IV curves for each of the devices in Fig. 1. IC-CAP, an HP software product for modeling semiconductor devices, was then used to extract a SPICE model for each device. Only the NMOS devices were actually simulated. The PMOS models were created from the NMOS models with appropriate modifications in mobility and series resistance to yield half the current drive of the corresponding NMOS. These device models were then used to simulate inverter chains as shown in Figs. 8 and 9. The load capacitance was varied to approximate fanouts of 3 and 7. Interconnect loading was ignored. The results are shown in Figs. 10 and 11. Fig. 10 shows that the gate delay improves about 30% per generation with the scaling described in the previous section. This is nearly the same as the historical trend of previous generations. Note that for L_g = 0.1 μm the gate delay (fanout = 1) is less than 15 ps. This is faster than the best that can currently be obtained with bipolar ECL. Fig. 11 shows the dependence of gate delay on the power supply. The stars denote the operating point from constant-field scaling. Note that these highly scaled devices offer high-speed operation even at low supply voltages. For example, the 0.1-μm generation should yield 23-ps gate delay (fanout = 1) even with V_{dd} = 0.5V. This would be excellent for low-power applications assuming that the high off-state leakage could be dealt with.

Off-State Leakage

The previous sections show that constant-field scaling of MOSFETs leads to a continuation of the historical trends of gate-level performance improvement. However, this comes at the price of exponentially increasing off-state leakage currents. For example, if an advanced circuit had 50 million micrometers of device width producing leakage current at 1 μA/μm, the quiescent supply current would be 50A. Clearly this is unacceptable. There are several proposals for dealing with this problem and I will briefly discuss some of them in this section. At this time we do not know the best way to deal with this problem.

One obvious solution to control quiescent power consumption is to put almost all the circuit in power-down mode at any instant and activate only those blocks that are being accessed. This system-level type of solution is beyond the scope of this paper and needs to be evaluated by the design community.

Another possible solution that has been proposed involves multiple threshold devices in the same technology. For example, the 0.1-μm generation could offer FETs with V_t = 0.2V and V_t = 0.4V. The low-V_t FETs could be used for speed-critical paths and the higher-V_t FETs could be used for tasks for which speed is not as important.

After modifying the doping profiles in TSUPREM-4 to get higher thresholds, the MEDICI simulations were repeated and new SPICE models extracted. Fig. 12 shows the resulting drive current and off-state current for various values of V_t in the 0.1-μm generation. Fig. 13 shows the gate delay as a function of V_t. From these graphs, FETs with V_t = 0.4V would yield gate delays about 80% longer than V_t = 0.2V but with off-state currents reduced by nearly three orders of magnitude. Again, the off-state currents shown are for nominal devices and worst-case would be higher. This approach is conceptually easy to implement in any technology. However, it increases the complexity of both the process and the circuit design.

Fully depleted (FD) silicon-on-insulator (SOI) devices have been proposed to reduce off-state current for a given V_t. These devices have a steeper subthreshold slope than conventional bulk devices, thus reducing off-state current without increasing V_t. However, single-gate FD SOI devices are difficult to scale into the deep submicrometer regime. Dual-gate FD SOI devices scale much better but are very complicated to make. These difficulties, coupled with the material quality and availability issues, make the FD SOI device an unlikely candidate for future generations of high-speed digital technology.

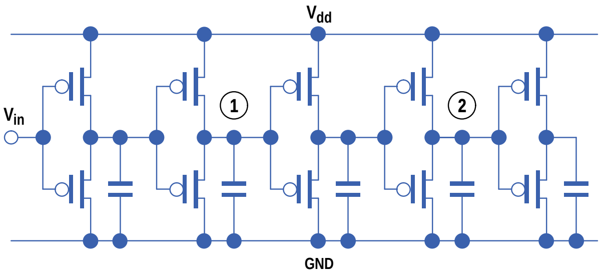

Fig. 8. *Inverter chain.*

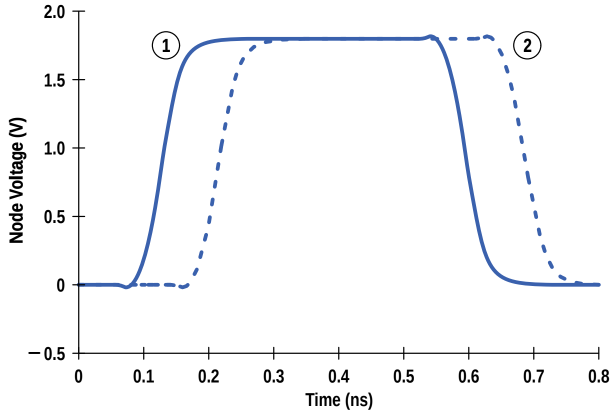

Fig. 9. *Inverter switching waveforms at nodes 1 and 2 of Fig. 8.*

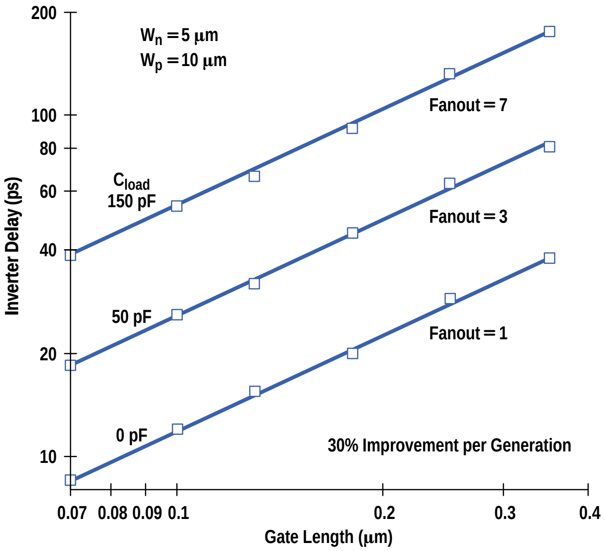

Fig. 10. *Inverter delay versus gate length.*

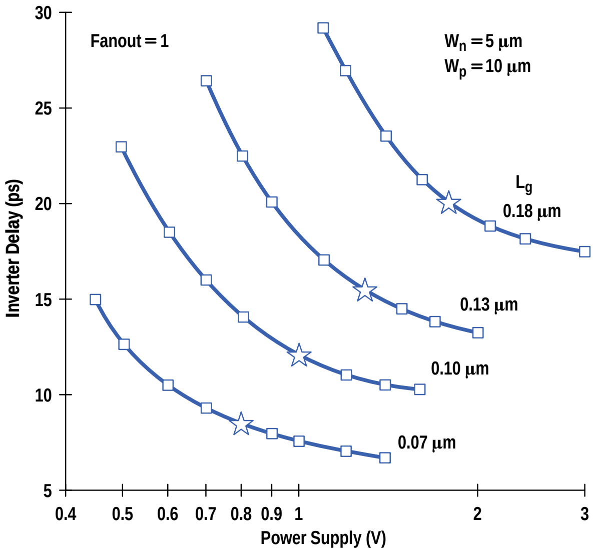

Fig. 11. *Inverter delay versus power supply voltage. The stars show the expected operating points. W_n and W_p are the widths of the n-channel and p-channel transistors.*

If no other solution for high I_{off} can be found, then V_t cannot be scaled lower than a certain point. For example, if one needed to keep I_{off} (nominal) at 1 nA/µm, then V_t (nominal) could not go below about 0.35V. We can apply this to the 0.1-µm generation (T_{ox} = 2.5 nm) and resimulate the device with V_t = 0.35V. After compact model extraction and inverter simulations, we find that V_{dd} must be increased to 1.8V to get the same performance as shown in Fig. 10 for the 0.1-µm generation. At V_{dd} = 1.8V and V_t = 0.35V, the device simulations predict a drive current of slightly over 1 mA/µm (NMOS). The peak oxide field would be over 7 MV/cm and the peak electron temperature would be about 3300K at $V_d = V_g$ = 1.8V (compare to Fig. 7). Even if we could obtain this very high drive current, it is questionable whether such a device could be created with adequate reliability. In any case, it is clear from this discussion that ceasing threshold voltage scaling would have a crucial impact on future device technologies.

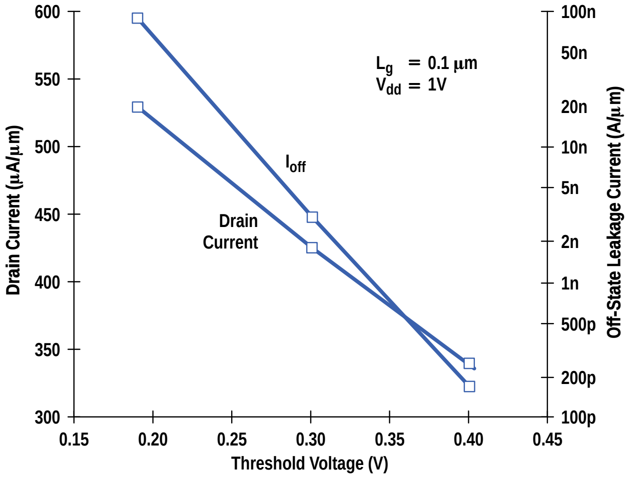

Fig. 12. *Drain current and off-state leakage current I_{off} versus threshold voltage V_t for the 0.1-μm generation.*

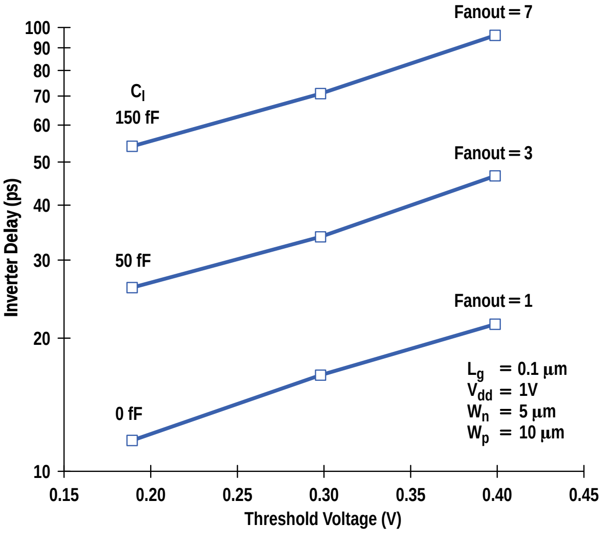

Fig. 13. *Inverter delay versus threshold voltage V_t for the 0.1-μm generation.*

Conclusions

We have explored MOSFET scaling into the future, extrapolating past scaling trends in channel length and gate oxide thickness. This scaling requires ever-shallower junction profiles and, below L_g = 0.15 μm, retrograde channel profiles. Constant-field scaling applied to V_{dd} and V_t continues the historical trend of about 30% improvement in gate delay per generation. In this era, MOSFET drive current remains nearly constant from one generation to the next and most of the performance improvement comes from the decreasing supply voltage. However, this performance comes at the price of exponentially increasing off-state leakage. Possible alternatives to this problem were discussed briefly but no clear resolution is available at this time. Clearly, this is an area where the design and technology communities must work together to develop an optimal roadmap for future device scaling.

▶ Go to Next Article
▶ Go to Journal Home Page