MOS Scaling: Transistor Challenges for the 21st Century

Scott Thompson, Portland Technology Development, Intel Corp.
Paul Packan, Technology Computer Aided Design, Intel Corp.
Mark Bohr, Portland Technology Development, Intel Corp.

Index words: SDE, transistor, scaling

Abstract

Conventional scaling of gate oxide thickness, source/drain extension (SDE), junction depths, and gate lengths have enabled MOS gate dimensions to be reduced from 10μm in the 1970's to a present day size of 0.1μm. To enable transistor scaling into the 21st century, new solutions such as high dielectric constant materials for gate insulation and shallow, ultra low resistivity junctions need to be developed. In this paper, for the first time, key scaling limits are quantified for MOS transistors (see Table 1). We show that traditional SiO_2 gate dielectrics will reach fundamental leakage limits, due to tunneling, for an effective electrical thickness below 2.3 nm. Experimental data and simulations are used to show that although conventional scaling of junction depths is still possible, increased resistance for junction depths below 30 nm results in performance degradation. Because of these limits, it will not be possible to further improve short channel effects. This will result in either unacceptable off-state leakage currents or strongly degraded device performance for gate lengths below 0.10μm. MOS transistor limits will be reached for 0.13μm process technologies in production during 2002. Because of these problems, new solutions will need to be developed for continued transistor scaling. We discuss some of the proposed solutions including high dielectric constant gate materials and alternate device architectures.

FEATURE	LIMIT	REASON
Oxide Thickness	2.3 nm	Leakage (I_{GATE})
Junction Depth	30 nm	Resistance (R_{SDE})
Channel Doping	V_T=0.25 V	Leakage (I_{OFF})
SDE Under Diffusion	15 nm	Resistance (R_{INV})
Channel Length	0.06μm	Leakage (I_{OFF})
Gate Length	0.10μm	Leakage (I_{OFF})

Table 1: Fundamental scaling limits for conventional MOS devices

Introduction

For more than 30 years, MOS device technologies have been improving at a dramatic rate [1,2]. A large part of the success of the MOS transistor is due to the fact that it can be scaled to increasingly smaller dimensions, which results in higher performance. The ability to improve performance consistently while decreasing power consumption has made CMOS architecture the dominant technology for integrated circuits. The scaling of the CMOS transistor has been the primary factor driving improvements in microprocessor performance. Transistor delay times have decreased by more than 30% per technology generation resulting in a doubling of microprocessor performance every two years. In order to maintain this rapid rate of improvement, aggressive engineering of the source/drain and well regions is required. In this paper, key methods for improving device performance are discussed. Creating shallow source/drain extension (SDE) profiles for improved short channel effects, the use of retrograde and halo well profiles to improve leakage characteristics, and the effect of scaling the gate oxide thickness are discussed in detail. Fundamental tradeoffs and scaling trends in engineering these effects are analyzed through experimental data and computer simulations. The impact of these trends associated with circuit requirements including power supply, threshold voltage, and off-state leakage on transistor design is also explored. We show that the scaling trends of the last ten years will be extremely difficult if not impossible to maintain unless new methods for device improvement are found. In addition to the conventional MOS transistor, several alternate device architectures are analyzed to understand the potential gains and tradeoffs associated with each device. The ability to overcome current physical technology limits such as gate oxide thickness and shallow junction formation as well as tradeoffs in circuit design will

determine if MOS transistors can be scaled into the next century.

Oxide Scaling

Gate oxide thickness scaling has been instrumental in controlling short channel effects as MOS gate dimensions have been reduced from 10µm to 0.1µm. Gate oxide thickness must be approximately linearly scaled with channel length to maintain the same amount of gate control over the channel to ensure good short channel behavior. Figure 1 plots the electrical channel length divided by gate oxide thickness for Intel's process technologies over the past 20 years. Each data point represents a process technology, developed approximately every three years, which was used to fabricate Intel's leading-edge microprocessors.

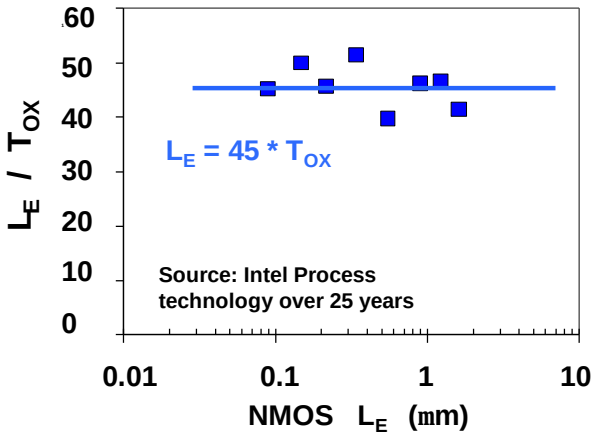

Figure 1: Channel length divided by gate oxide thickness versus channel length

From Figure 1, a simple relationship between oxide thickness and the minimum channel length set by short channel effects is observed:

$$L_E = 45 * T_{OX} \qquad \text{(Eq. 1)}$$

This relationship exists because the channel depletion layer is engineered to become smaller as the gate oxide thickness is decreased. In addition, short channel behavior is governed by the ratio of channel depletion layer thickness to channel length. The channel depletion layer is inversely proportional to the square root of the channel doping concentration. During device optimization, channel doping is increased as the oxide is scaled to maintain approximately the same device threshold voltage. Figure 2 illustrates this point. In Figure 2, the thickness of the channel depletion layer for two devices with different oxide thicknesses is shown. Figure 2a shows the depletion layer for a device with an oxide thickness of 4.5 nm while Figure 2b shows a device with an oxide thickness of 3.2 nm.

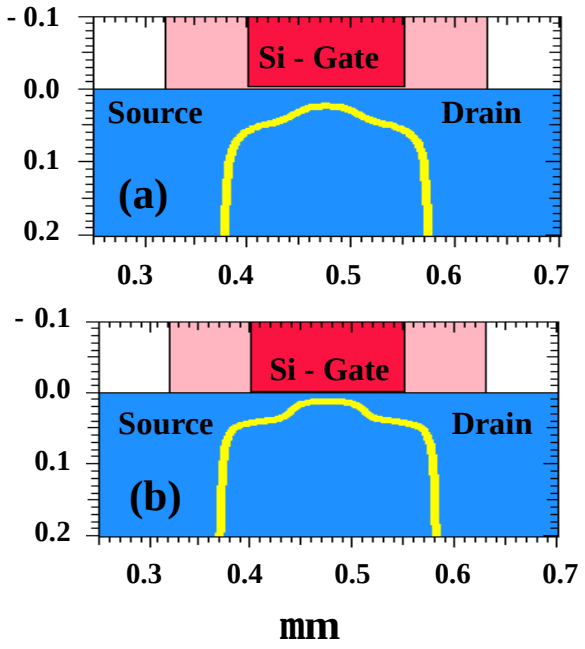

Figure 2a and 2b: Device simulations showing channel depletion layer thickness for devices with two oxide thicknesses: (a) 4.5 nm, (b) 3.2 nm

Both devices have the same off-state leakage. The device with the thinner oxide has a smaller channel depletion layer and hence improved short channel characteristics. The improved short channel effects can be taken advantage of by targeting a smaller channel length. Thus, for continued MOS channel length scaling, the gate dielectric thickness must continue to be scaled. Figure 3 shows the Semiconductor Industry Association's (SIA) road map for gate dielectric thickness. This roadmap predicts that continued gate dielectric scaling will be required with a new gate dielectric material needed for the 2002-2005 time frame.

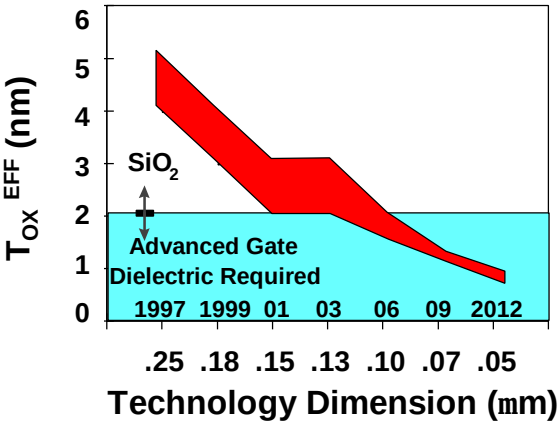

Figure 3: SIA road map for junction depth

Scaling Limit for SiO_2

SiO_2 or nitrided SiO_2 has been the gate dielectric used by the semiconductor industry for over 30 years. The thickness limit is the same for both materials and is not limited by manufacturing control. Today, it is technically feasible to manufacture 1.5 nm and thinner oxides on 200 mm wafers [3]. The thickness limit for SiO_2 is set instead by gate-to-channel tunneling leakage. Figure 4 schematically shows the tunneling leakage process for an NMOS device biased in inversion.

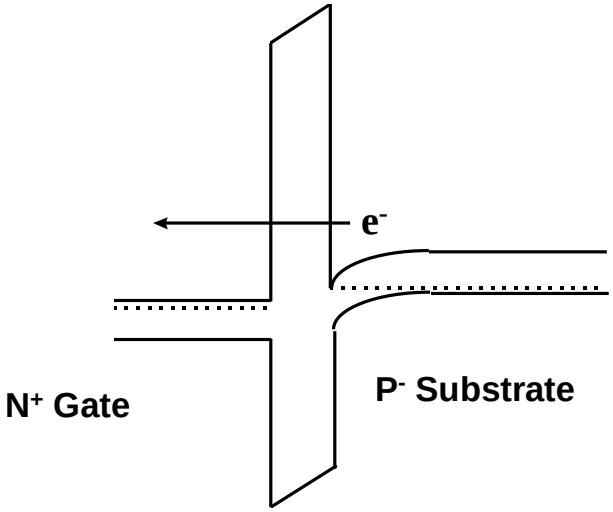

Figure 4: Direct tunneling leakage mechanism for thin SiO_2

As the thickness of the dielectric material decreases, direct tunneling of carriers through the potential barrier can occur. Because of the differences in height of barriers for electrons and holes, and because holes have a much lower tunneling probability in oxide than electrons, the tunneling leakage limit will be reached earlier for NMOS than PMOS devices. The SiO_2 thickness limit will be reached approximately when the gate to channel tunneling current becomes equal to the off-state source to drain sub-threshold leakage (currently ~1nA/µm). Figure 5 shows the area component of gate leakage current in A/cm^2 versus gate voltage. If we assume the gate leakage limit occurs for devices with 0.1µm gate length designed for 1.0V operation, the SiO_2 thickness limit occurs at ~1.6 nm.

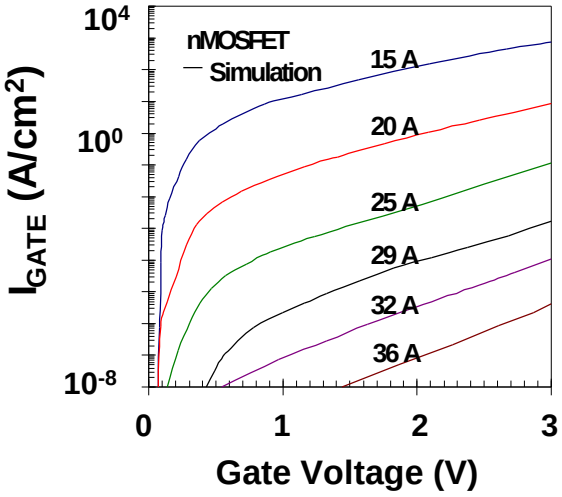

Figure 5: Gate leakage versus gate voltage for various oxide thicknesses [5]

We now have established that the thickness limit for SiO_2 is ~1.6 nm. However, due to quantum mechanical and poly-Si gate depletion effects, both the gate charge and inversion layer charge will be located at a finite distance from the SiO_2/Si interfaces with the charge location being a strong function of the bias applied to the gate. Figure 6 shows the location of the inversion layer charge in the silicon substrate for a transistor with a typical bias when quantum mechanical effects are taken into account [4]. The centroid for the inversion charge is ~1.0 nm from the SiO_2/Si interface. This increases the effective SiO_2 thickness (T_{OX}^{EFF}) by ~0.3 nm. By taking into account the charge distribution on both sides of the gate, the minimum effective oxide thickness for a MOS device bias in inversion (at voltages used in our 0.25 or 0.18µm technologies) is increased by approximately 0.7 nm. Thus, the 1.6 nm oxide tunneling limit results in an effective oxide thickness of approximately 2.3 nm.

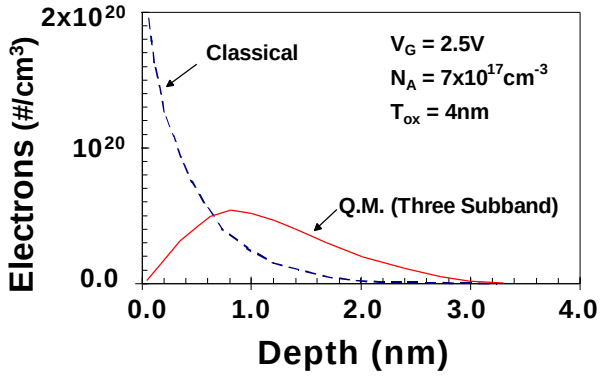

Figure 6: Position of inversion channel charge versus depth

Based on the previous arguments for controlling short channel effects, a limit for SiO_2 thickness will set a limit on the gate and channel length of MOS devices. Figure 7 plots gate and channel length versus effective oxide thickness. From this figure, we see that the limit for gate and channel length for an SiO_2 gate dielectric MOSFET is 0.1µm and 0.06µm, respectively. Since in leading-edge logic technologies, the gate dimension is printed smaller than the technology features, the SiO_2 thickness limit and the gate length limit will be reached for ~0.13µm technologies.

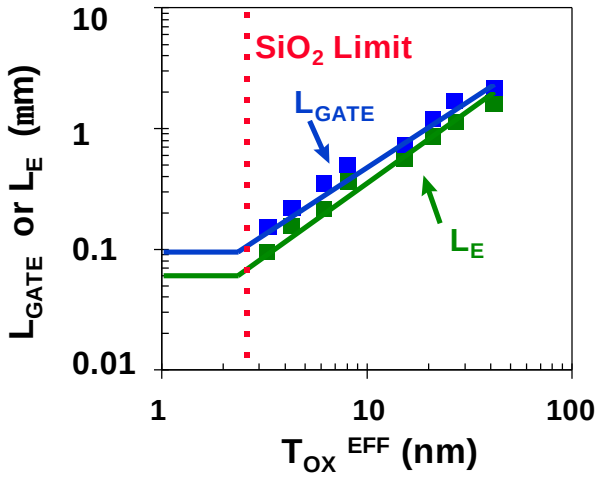

Figure 7: Gate and channel length versus effective oxide thickness

Alternative High Dielectric Constant Materials

Alternative high dielectric constant materials will be the key to continued MOSFET scaling past 0.1µm gate dimensions. With these materials, thicker dielectric layers can be used yet the same inversion layer characteristics can be maintained. These thicker layers result in less carrier tunneling, and they permit further scaling of the effective oxide thickness. Table 2 lists the leading alternative dielectrics and their status.

OPTION	ISSUES / STATUS
Si_3N_4 /	Small advantage especially with buffer layer
nitride	Close to being ready (G. Lucovsky, T. P. Ma)
Ta_2O_5	Need SiO_2 buffer/ no poly-silicon gate
	Very early stages (S. Kamiyama)
TiO_2	Need SiO_2 buffer/ no poly-silicon gate
	Very early stages (S. A. Campbell)
BST	Deep states/ buffer layer/ no poly-silicon gate
	Early stages FET (large DRAM interest)

Table 2: Alternate high dielectric constant materials [6-9]

All these materials, with the possible exception of Si_3N_4, need an SiO_2 buffer layer between the high dielectric constant materials and the silicon substrate in order to obtain an interface with low interface states. They also need a metal electrode to eliminate a reaction between the alternate dielectric and the poly-Si that usually forms SiO_2. This is extremely unfortunate since it can be shown that if an SiO_2 buffer layer is needed, and since quantum mechanical effects and poly-Si gate depletion cannot be eliminated, an Si_3N_4 gate dielectric with a buffer layer can only improve the effective oxide thickness by 0.3 nm before it reaches its tunneling thickness limit [10]. The problem with using a metal gate electrode with an alternative dielectric material is that the metal gate is not compatible with deep sub-micron complementary CMOS devices. A metal gate with a work function equal to intrinsic silicon such as tungsten would produce complementary CMOS devices. However, a mid-bandgap gate metal is not compatible with deep sub-micron devices because of degraded short channel behavior. Figure 8 shows the depletion layer obtained from a device simulator for two NMOS devices with the same threshold voltage but with different gate electrodes: (a) with an N+ poly-Si gate and (b) with a tungsten gate. As can be seen from this figure, the device with the tungsten gate has a significantly larger depletion layer and hence degraded short channel behavior.

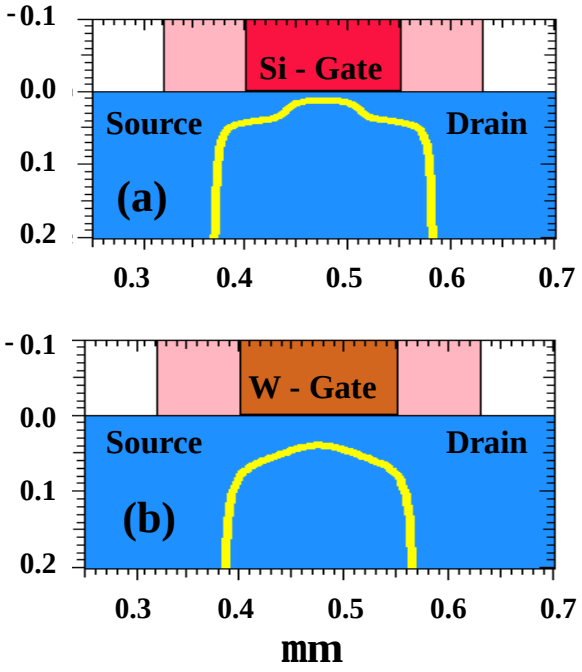

Figure 8: Device simulation of two devices showing depletion layers: a) N+ poly-Si and b) tungsten gate

Source/Drain Engineering

In this section, we investigate the scaling of source/drain extension (SDE) depth and gate overlap for MOSFETs of 0.1µm and below. For the purposes of this discussion, the SDE is the shallow diffusion that connects the channel with the deep source and drain. Junction depth always refers to the SDE junction depth. The deep source/drain junction depth is held constant. Overlap is defined as the distance the SDE extends under the gate. The metallurgical spacing (L_{MET}) is the distance between the source and drain SDE (see Figure 9).

We show that a minimum SDE to gate overlap of 15-20 nm is needed to prevent degradation of drive current (I_{DSAT}). We also show that scaling SDE vertical depths below 30-40 nm results in little to no performance benefit for 0.1µm devices and beyond. This is because any improvement in short channel effects due to reduced charge sharing is offset by a large increase in external resistance and too small an overlap between the SDE and gate.

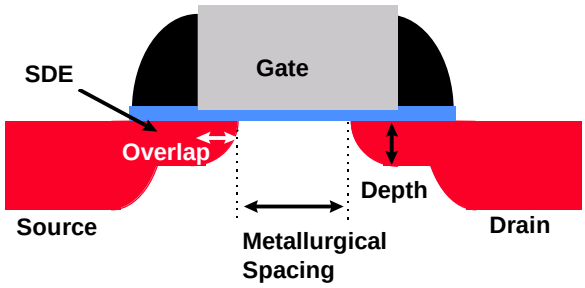

Figure 9: Terminology used in this discussion

Shallow Junction Formation

Very short gate length transistors with shallow SDE junctions and small gate overlap have been reported [11,12]. Many of these transistors have lower than expected drive currents given their extremely short channel lengths. We propose that these low drive currents are the result of an SDE that is too shallow and therefore leads to a high external resistance and too small of an overlap between the SDE and gate. Junction depths are currently 50-100 nm for 0.25µm process technologies and are predicted to be as low as 10 nm for future deep sub-micron devices (see Figure 10). The fabrication of these shallow junctions is less of an issue than whether or not the shallow junctions offer any device benefit. Shallow junctions can be fabricated by carefully controlling transient enhance diffusion (TED) [13-17]. Methods for reducing TED include lowering implant energies, amorphization followed by solid phase expitaxial regrowth and high temperature, and short time rapid thermal anneal cycles. Figure 11 shows an example of a shallow 35 nm junction formed by a low energy implant and a rapid thermal anneal. Alternate architectures such as removable spacer process flows can also be used to minimize SDE depths. In this architecture, an initial disposable spacer is used. High temperature cycles for forming the S/D and doping the poly-Si gate are used before the introduction of the SDE structure. These cycles permit the use of extremely low temperature anneal cycles engineered to minimize SDE junction depths and maximize dopant concentrations.

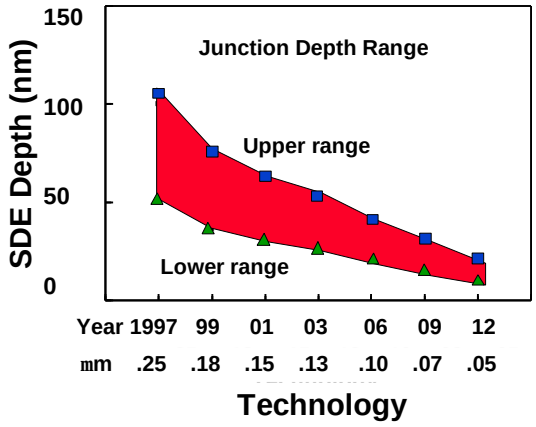

Figure 10: SIA roadmap for junction depth

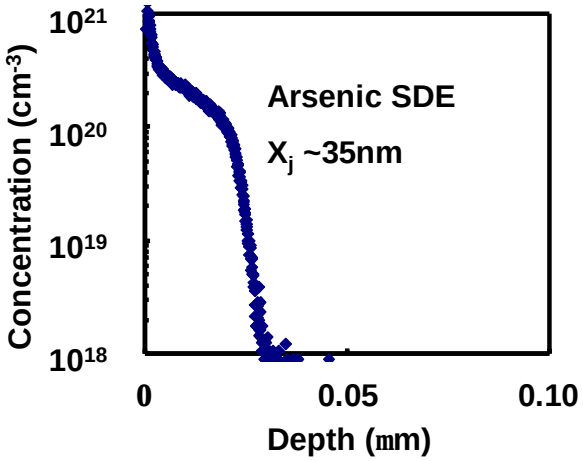

Figure 11: Shallow 30.0 nm SDE formed by a low energy implant and rapid thermal anneal

SDE Junction Scaling

Reducing SDE junction depths will improve device short channel characteristics by reducing the amount of channel charge controlled by the drain. This may not,

however, lead to improved device performance. Figures 12a and 12b show the potential contours for two devices with junction depths of 30 and 150 nm, respectively, biased in the off-state condition. In this figure, the potential contours extend much further into the channel for the device with the deep junction.

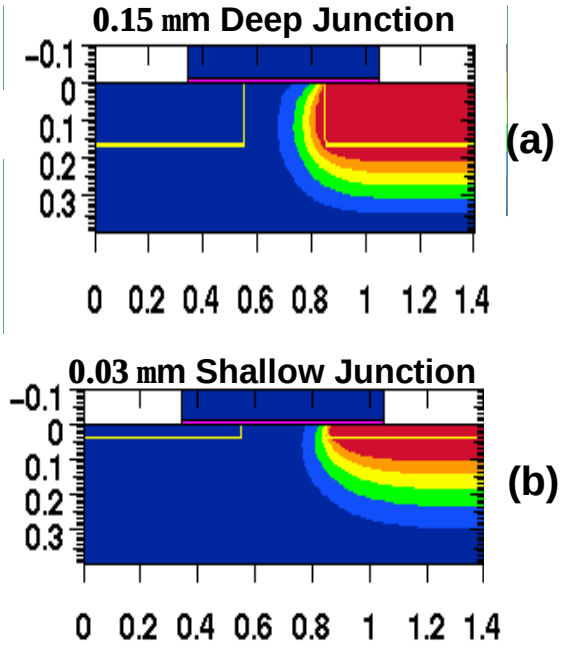

Figure 12: Potential contours for two devices biased in an off-state condition (a) 30 nm shallow junction and (b) 150 nm deep junction

Thus, transistors with deeper junctions will have worse short channel characteristics. Unfortunately, shallow SDE junctions can increase the external resistance of the device. Figure 13 shows the various components of external resistance for a MOS device. Current flows from the channel inversion layer into the SDE accumulation region ($R_{ACCUMULATION}$). The current then spreads out into the SDE ($R_{SPREADING}$) region and through the bulk SDE area (R_{SHUNT}). The final component of resistance is associated with the deep source/drain and salicide ($R_{CONTACT}$). In deep sub-micron devices, particularly NMOS, the SDE accumulation and spreading components are the dominant components of external resistance. The components associated with the SDE region become a greater problem as the transistor feature size is scaled (channel length and SDE depth reduced) since the scaling reduces channel resistance while increasing the components of SDE resistance.

A second scaling limit is the minimum SDE-to-gate overlap for a device. Reducing this overlap causes the current to spread out into a lower doping location of the SDE. This can strongly increase accumulation and spreading resistance and increase the total external resistance. For example, if the overlap is zero, the current flow would spread out at the gate edge where the SDE doping concentration would be zero. In the next section, we investigate scaling limits for SDE to junction depth and gate overlap.

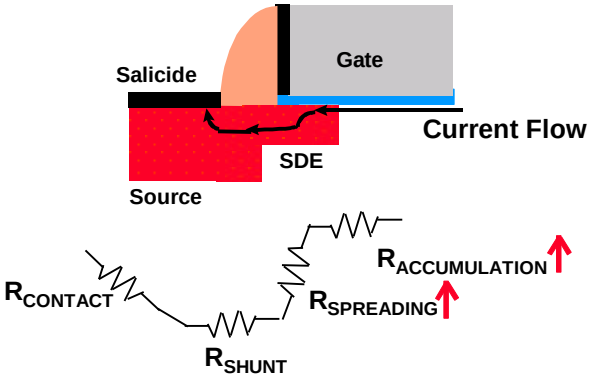

Figure 13: Components of external resistance

Minimum SDE-to-Gate Overlap

The test structure shown in Figure 14 is used to evaluate the effect of SDE-to-gate overlap on I_{DSAT}. In this test structure, the SDE implant is performed after the formation of a thin offset spacer. By varying the thickness of the offset spacer, the SDE-to-gate overlap and vertical junction depth can be independently varied. The transistor data presented are measured on devices with a process flow similar to our 0.25µm technology [2].

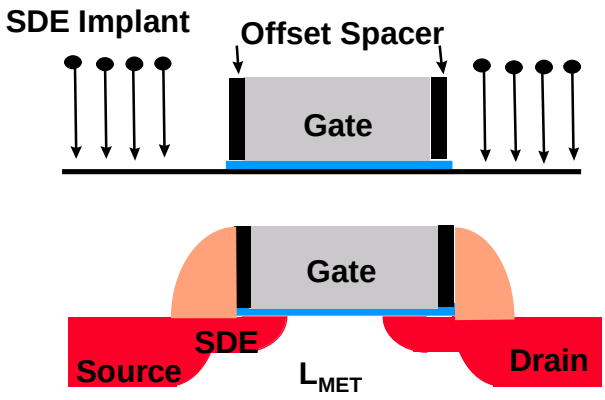

Figure 14: Test structure to evaluate minimum SDE-to-gate overlap

Also included is data on transistors with gate length, gate oxide, and power supply scaled by 0.7 and $(0.7)^2$ from our 0.25µm technology. All transistors have controlled sub-threshold slopes of less than 85mV/decade, 1nA/µm off-state leakage, and electrical channel lengths (L_E) between 0.06 and 0.14µm.

With the above test structure fabricated for a range of poly-Si gate lengths, the transistor saturation drive current versus the SDE overlap for both fixed vertical SDE depth and fixed SDE metallurgical spacing was measured. The SDE metallurgical spacing is kept constant by adjusting the poly-Si gate length to maintain 1nA/µm off-state leakage. Figure 15 shows the vertical SIMS profile of an SDE junction used in the experiment (1.0e15cm^{-2}, 5keV arsenic implant RTA annealed). Figure 16 shows the effect of spacer offset on overlap capacitance and I_{DSAT}. For spacer offsets greater than 40 nm, there is a flattening in overlap capacitance implying minimal SDE-to-gate overlap. A degradation in I_{DSAT} is also clearly observed for offset spacer widths greater than 20 nm.

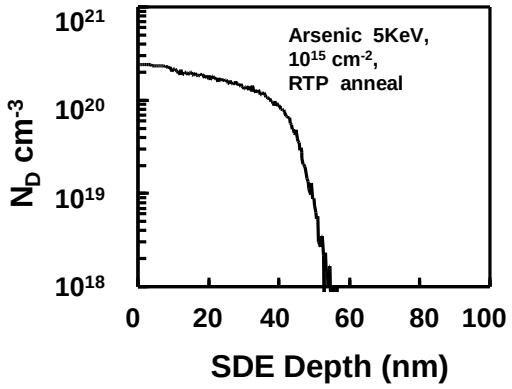

Figure 15: Vertical SIMS profile of Arsenic SDE

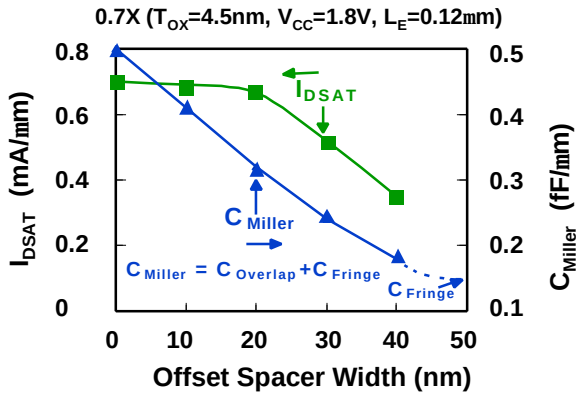

Figure 16: I_{DSAT} and C_{MILLER} versus spacer offset

The lateral diffusion of the SDE junction under the gate edge is estimated to be 0.6 - 0.7 times the vertical depth minus the offset spacer width. This estimate is obtained from process simulations and junction-staining measurements. Experimentally, the offset spacer width is varied from 0 to 40 nm and is used to modulate the SDE-to-gate overlap from approximately 40 to 0 nm. Figures 17 and 18 show I_{DSAT} versus SDE overlap for both NMOS and PMOS 0.25µm devices as well as the 0.7 scaled devices. These figures also show that, independent of the feature size of the process technology, a degradation in I_{DSAT} is observed if the overlap is less than 15-20 nm.

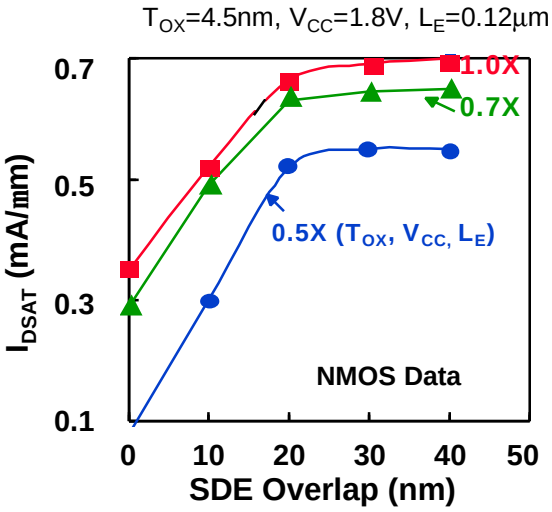

Figure 17: I_{DSAT} versus SDE overlap (NMOS)

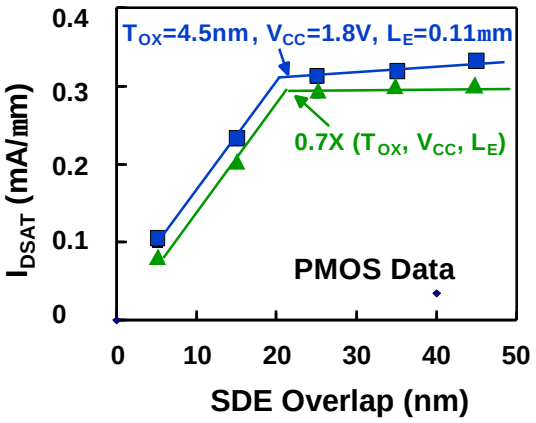

Figure 18: I_{DSAT} versus SDE overlap (PMOS)

Minimum SDE Junction Depth

The optimal SDE vertical depth is now investigated. For this set of experiments, both the conventional and removable spacer flows were used. Figure 19 shows NMOS and PMOS drive current versus SDE depth for devices with 1nA/µm of off-state leakage. The SDE depths were adjusted by varying the implant energy (500eV - 40KeV) and the RTA temperature. In Figure 19, we see that a maximum in I_{DSAT} occurs when the vertical junction depth is 35-40 nm. With an SDE deeper than 35-40 nm, short channel effects degrade due to increased charge sharing. This necessitates a larger channel length to meet the off-state criteria and a loss in I_{DSAT}. SDE depths shallower than 35-40 nm result in degraded I_{DSAT} due to increased external resistance and an overlap between the SDE and gate that is too small.

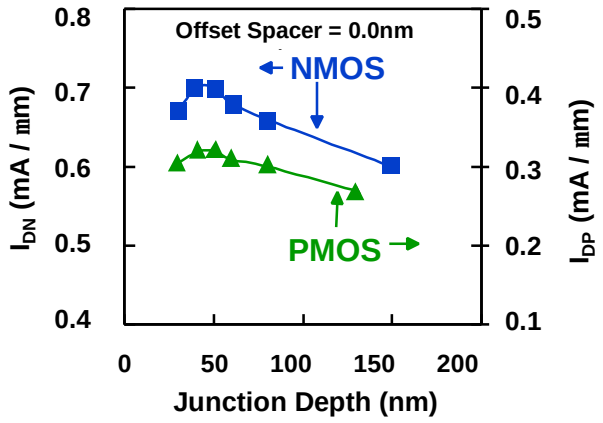

Figure 19: I_{DSAT} versus SDE depth

Simulation results for the above experiment are shown in Figure 20. In this figure, external resistance and short channel behavior (defined by source-to-drain distance at 1nA/µm off-state leakage) versus SDE junction depth are quantified.

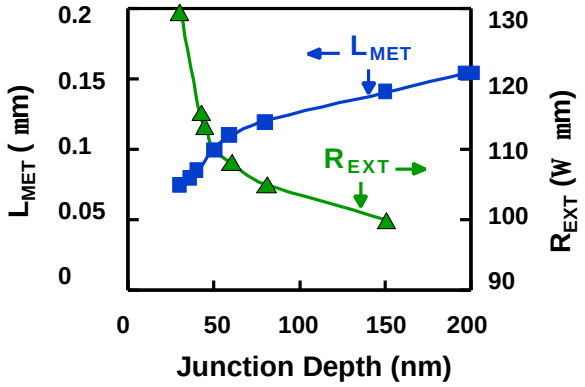

Figure 20: Simulation data quantifying R_{EXT} and L_{MET} versus junction depth

These results support the conclusion that the observed drive current maximum at a 35-40 nm junction depth results from tradeoffs in short channel effects, external resistance, and SDE-to-gate coupling. Note that these conclusions implicitly assume that the maximum SDE concentration is solid solubility limited for these devices.

Channel Engineering

Up to now we have shown how gate oxide thickness and junction scaling has enabled channel length scaling by improving short channel characteristics. We have also quantified scaling limits for these two techniques. The third and final technique to improve short channel characteristics is well engineering. By changing the doping profile in the channel region, the distribution of the electric field and potential contours can be changed. The goal is to optimize the channel profile to minimize the off-state leakage while maximizing the linear and saturated drive currents. Super Steep Retrograde Wells (SSRW) and halo implants have been used as a means to scale the channel length and increase the transistor drive current without causing an increase in the off-state leakage current [18-23]. Figure 21 is a schematic representation of the transistor regions that are affected by the different types of well engineering. Retrograde well engineering changes the 1D characteristics of the well profile by creating a retrograde profile toward the Si/SiO_2 surface. The halo architecture creates a localized 2D dopant distribution near the S/D extension regions. The use of these two techniques to increase device performance is discussed in the following sections. We show that channel doping optimization can improve

circuit gate delay by ~10% for a given technology. However, we also show that well doping engineering cannot provide the generation after generation channel length scaling that gate oxide and SDE junction depth scaling has provided.

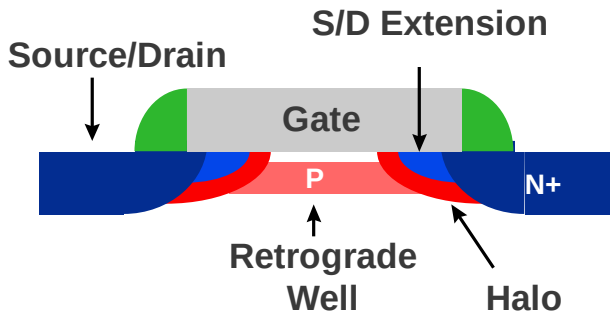

Figure 21: Schematic representation of different aspects of well engineering

Retrograde Well Engineering

The use of retrograde well profiles to improve device performance has been reported [18,21]. The retrograde profile is typically created by using a slow diffusing dopant species such as arsenic or antimony for PMOS devices and indium for NMOS devices. It has been established that SSRW can improve short channel effects, increase surface mobility, and can lead to either an increase or a decrease in saturated drive current depending on a variety of technology issues [18-20]. Although retrograde wells do not appreciably improve saturated drive currents, we will show that for today's deep sub-micron technologies, they do improve linear drive currents and lead to improved circuit performance. Unfortunately, as S/D junction depths continue to decrease, this gain in linear drive current is further diminished.

The process flow used for the devices in this study has been reported [1]. In this study, aggressive SSRW wells created by indium (NMOS) and arsenic (PMOS) implants are compared to uniform wells formed by boron (NMOS) and phosphorus (PMOS). Figure 22 shows the vertical doping profile for an SSRW formed by an arsenic implant and by a conventional flat phosphorus well. As can be seen, the well doping profile formed by the arsenic implant is clearly retrograde to the surface. Although the SSRW profile has a lower surface concentration, the profile was engineered to give the same threshold voltage as the flat well case to ensure an accurate comparison.

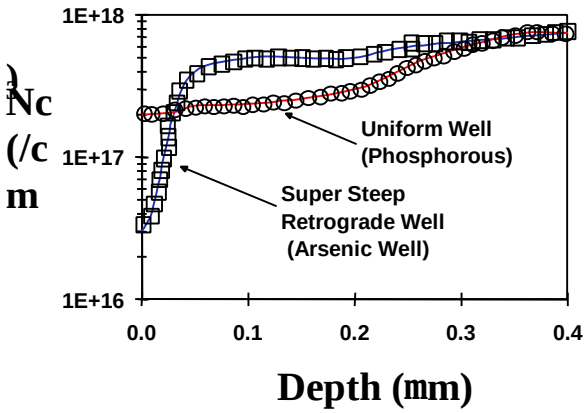

Figure 22: Vertical concentration doping profile for SSRW and conventional well doping profiles

Figure 23 shows the minimum channel length that can be supported for an off-state leakage current of 1nA/µm for a range of threshold voltages for both SSRW and uniform well transistors. As expected, higher threshold voltages support smaller gate lengths due to the increase in channel doping. This figure shows that the SSRW architecture supports smaller channel lengths compared to the uniform well case for all threshold voltages. Similar results are seen for antimony (PMOS) and indium (NMOS). For the purposes of this paper, only PMOS data will be shown. Figures 24 and 25 compare I_{OFF} and I_{DSAT} versus electrical channel length for SSRW and uniform well transistors. Figure 24 shows improved source-to-drain leakage for the SSRW device for sub-0.25µm channel lengths implying improved short channel effects. However, Figure 25 shows a decrease in saturated drive current for the same SSRW device. Figure 26 shows families of curves for drain current versus drain voltage for SSRW and uniform well devices. The devices have a channel length of 0.15µm. For devices with the same channel length, the linear drive current is approximately equal, indicating no change in mobility for SSRWs. However, the current does saturate at a lower drain bias.

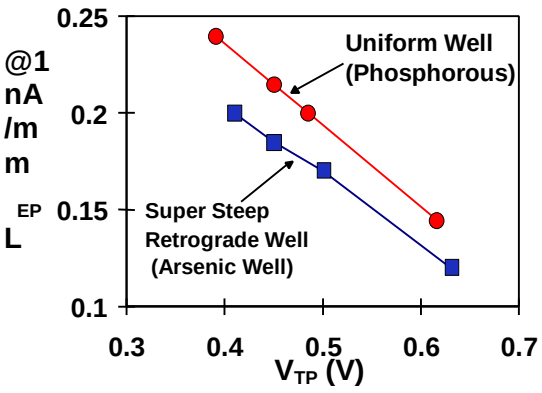

Figure 23: Channel length at which 1nA/µm of off-state leakage current occurs as a function of threshold voltage for SSRW and uniform well profiles

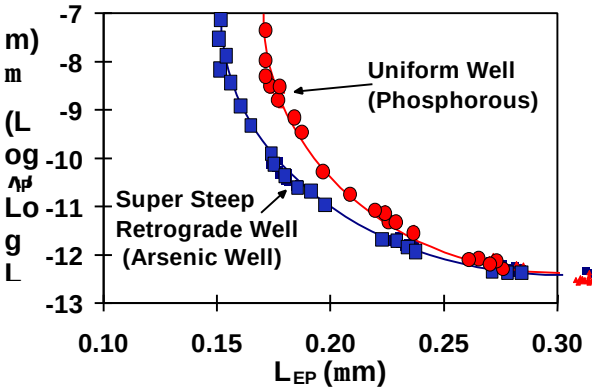

Figure 24: Leakage current as a function of channel length for SSRW and uniform well transistors with the same threshold voltage

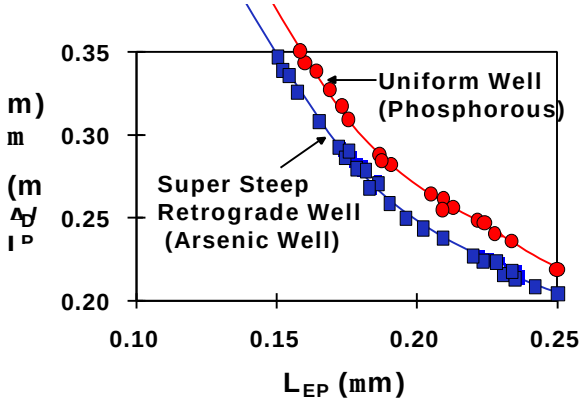

Figure 25: Saturated drive current (IDSAT) versus channel length for SSRW and uniform well transistors

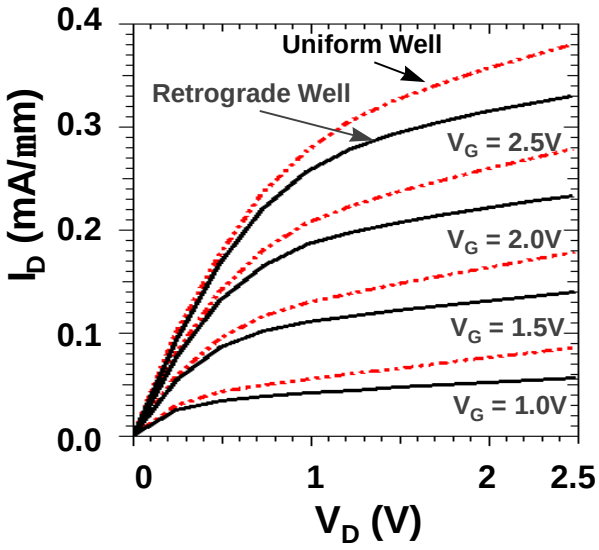

Figure 26: $I_D V_D$ characteristics for SSRW and uniform well devices as a function of gate voltage

In the next section, device simulations are used to understand this decrease in V_{DSAT}. Figure 27 shows the IV characteristics for SSRW and uniform well devices in which both devices have the same value of I_{OFF} (1nA/µm). Even though the SSRW device can support smaller channel lengths due to improved short channel effects, only a slight gain in I_{DSAT} is seen. The linear drive current, however, is clearly increased. For logic gate delays with fast input rise times and large loads, drive current in the linear mode is at least as important as drive current in saturation. Measured circuits showed that the increase in linear drive current improved inverter switching delays by up to 10%.

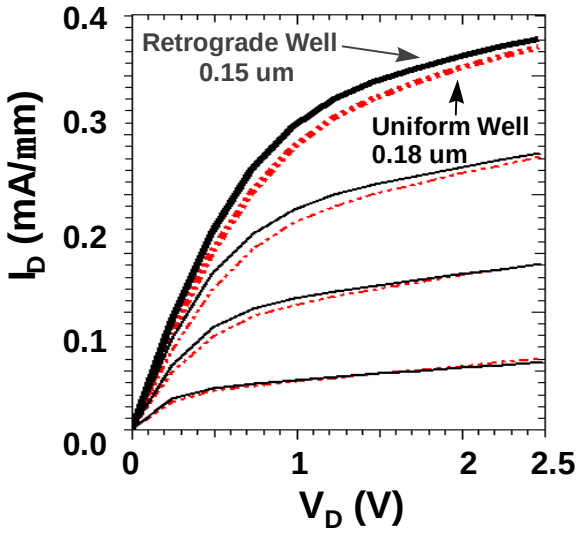

Figure 27: I_DV_D characteristics for SSWR and uniform well devices both having the same I_{OFF} criteria

Fundamental Operation of SSRW

In the classical derivation of the NMOS transistor, the drive current is calculated by integrating the inversion charge along the channel [24]:

$$I_D = \frac{W}{L}\int_{VS}^{VD} m_n\, Q_n(V)\, dV \qquad \text{Eq. 1}$$

It is typically assumed that the depletion charge and V_T are constant along the channel for this calculation. As shown schematically in Figure 28, the depletion charge and V_T actually increase along the channel from source to drain due to the body effect. This is true for both the SSRW and uniform well device. However, the increase in depletion charge and consequently V_T is larger for the SSRW device because of the higher doping in the substrate (see Figure 22) resulting in a larger body effect. The larger V_T for the SSRW device at high drain bias lowers the saturation voltage ($V_{DSAT}=V_G-V_{T(Drain)}$). This causes the reduction in I_{DSAT} for the SSRW device shown in Figure 26. The improvement in transistor performance due to SSRW strongly depends on the ability to scale the channel length due to improved short channel effects. Figure 29 shows the net change in performance due to SSRW versus junction depth. As S/D junction depths are scaled, the improvement in short channel effects from the use of SSRW decreases.

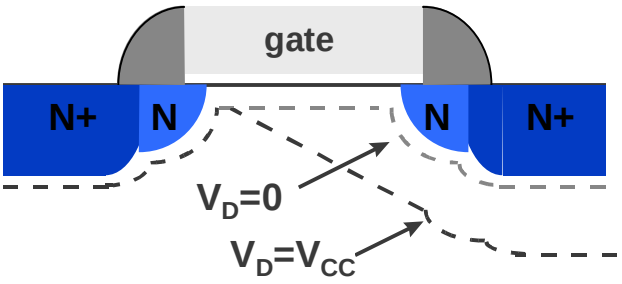

Figure 28: Schematic representation of the depletion layer for low and high drain bias

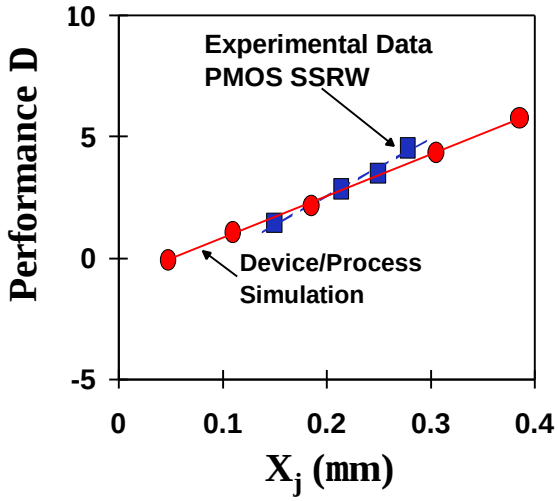

Figure 29: Improvement in device performance for SSRW over uniform well devices versus S/D depth

Halo Engineering

The addition of well implants to create a non-uniform well profile to improve short channel effects has been reported [25-27]. These implants may be vertical or angled and are typically done after gate patterning. They add additional well dopant around the source and drain regions providing an increased source-to-drain barrier for current flow. For long channel devices, the additional halo dopants only modestly change the threshold voltage. For short channel devices, however, a large increase in threshold voltage is seen. In order to maintain a constant threshold voltage for the target devices, the nominal threshold implant must be lowered for the halo devices (see Figure 30). This results in a lower long channel threshold voltage, and it can create a curvature reversal in the threshold voltage versus channel length curve. It will be shown in the following sections that although the use of halos can improve performance by compensating

for manufacturing variability, halos do not fundamentally improve device performance.

The process flow for the devices reported here has been presented previously [1,2]. Figure 30 shows a lateral surface cut of the doping profile for both a conventional and halo device. For the halo device, there is a lateral decay of the well doping profile toward the center of the channel. As the gate length of the halo device is decreased, the average well concentration increases resulting in a higher V_T.

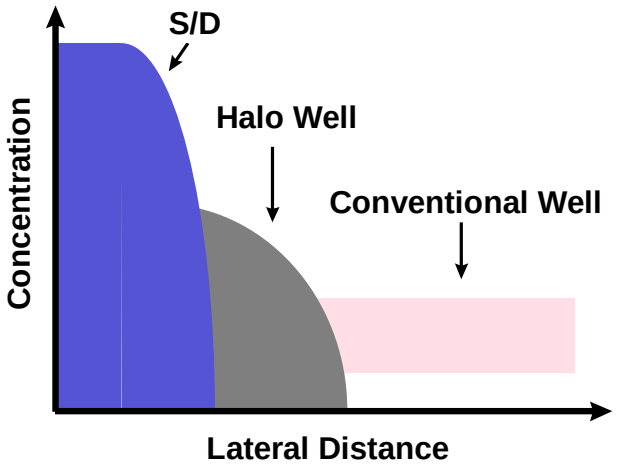

Figure 30: Schematic showing a lateral surface cut of the well doping near the Si/SiO_2 interface

Figures 31 and 32 show the threshold and off-state leakage characteristics versus channel length for conventional and halo devices. It should be noted that the change in well doping as a function of size makes extraction of effective channel length a strong function of extraction methodology for halo devices and often becomes much less meaningful. Because of this, it is often clearer to use I_{DSAT} versus I_{OFF} when comparing device performance for halo devices. Figure 33 shows I_{DSAT} versus I_{OFF} characteristics for a halo and non-halo device. As seen, there is very little improvement in I_{DSAT} at the targeted I_{OFF} for the halo device (Figure 33).

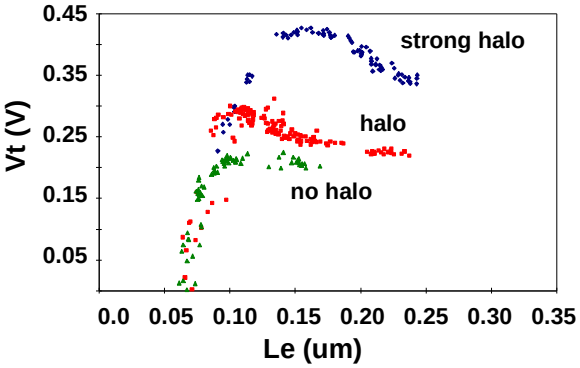

Figure 31: Threshold voltage as a function of channel length for a no halo, moderate halo, and strong halo device

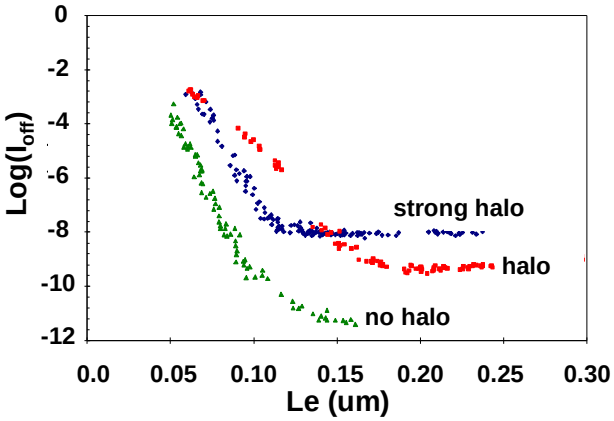

Figure 32: Off-state leakage current as a function of channel length for a no halo, moderate halo, and strong halo device.

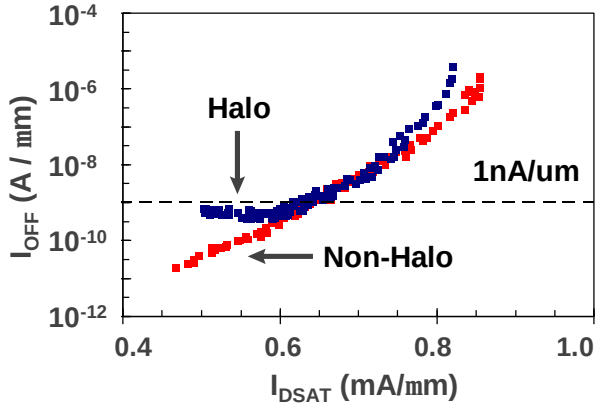

Figure 33: I_{OFF} versus I_{DSAT} for a halo and a conventional device (little to no gain in I_{DSAT} is seen for a given I_{OFF}.)

Fundamental Operation of Halo Well Profiles

Halo profiles are created by implanting extra dopants into the wells immediately after tip implantation. The implant is typically performed at an angle and energy high enough to ensure the implant dose is outside the final SDE profile. After spacer processing and S/D anneal, the resulting profile diffuses due to TED effects, resulting in a relatively flat profile over the dimensions of current device sizes. Figure 34 shows experimental results for the as implanted and final doping profile for a typical boron halo implant. The data includes the effect from damage generated by the SDE and S/D implants. As can be seen, the profile is quite flat over the characteristic channel length dimensions for today's 0.25µm and 0.18µm technologies. However, even though the halo profile is relatively flat, it still causes an increase in well doping as the gate length is decreased. This is because the same halo implant dose is confined in a smaller area. For flat well devices, I_{OFF} quickly decreases as the channel length is increased. This is due to the exponential relationship between the current and the potential barrier in the sub-threshold region. For the halo cases, the leakage current does not decrease as quickly with size. In fact, for extremely strong halos, an increase in I_{OFF} with increasing size can be seen. This can be explained by the change in the source-to-drain potential barrier for different size devices in the case of the halo well. For the strong halo devices, the threshold voltage is rapidly decreasing as the device size increases.

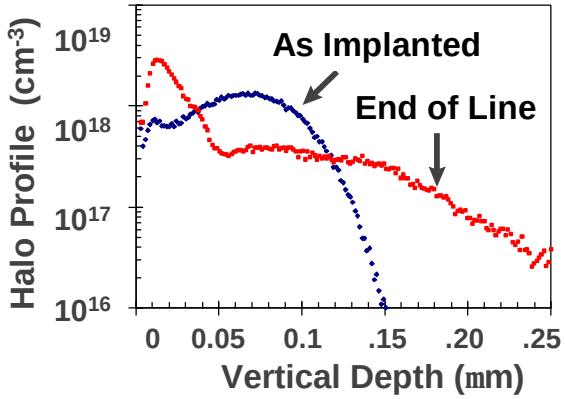

Figure 34: As implanted and end of line vertical halo profile (due to TED effects, a large amount of diffusion is seen)

This decrease compensates for the reduction in the electric field due to the increased channel length that results in less change in I_{OFF}. The strength of the halo depends not only on the halo doping concentration, but also on the lateral confinement of the halo. Figure 35 shows the simulation results on the effect of halo confinement for I_{OFF} versus device size. In this figure, I_{OFF} is plotted versus L_E for several values of **s** where **s** is defined as the characteristic lateral decay length of a gaussian halo doping profile, which begins at the transistor gate edge. Increasing the halo confinement increases the localization of the halo effect. A comparison of simulation and experimental results (Figures 32 and 35) shows that a relatively non-localized halo profile matches the experimental data. This is in agreement with the SIMS data of Figure 34. Therefore, for a single device size, both the halo and conventional device will have close to the same doping profile for the same off-state leakage criteria. However, there will be a large difference in the well doping level and threshold voltage for the device variations around this device. For the halo device, the threshold voltage will be lower for larger device sizes. Due to manufacturing variation, the target device will be necessarily larger than the worst-case device defined by maximum tolerable I_{OFF}. The gate drive (V_{CC}-V_T) for the target device is increased for the halo device resulting in an increase in I_{DSAT}. A halo can cause a greater than 10% increase in I_{DSAT} for the target device, relative to a non-halo process.

In order to scale deep sub-micron devices, halo implants must be used to improve the performance of target devices. Current technologies have used halo architectures to increase performance by up to 10%. Due to strong TED effects, halo profiles are not well confined in the technology now being used. A complicated interaction between halo dopant profiles, short channel effects, off-state leakage currents, and threshold voltages determines the final device performance gain.

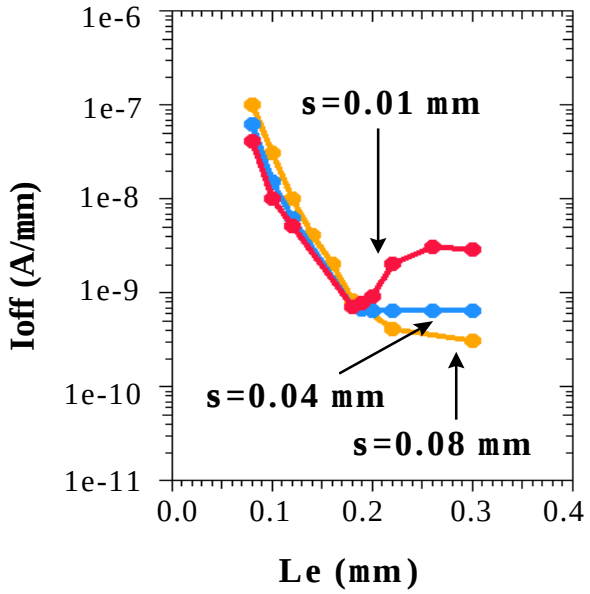

Figure 35: Simulation results showing the effect of halo confinement on I_{OFF} where σ is defined as the characteristic lateral decay length of a gaussian halo profile and is in units of µm.

The halo architecture does not improve device performance for the worst-case device, but instead provides a subtle benefit by improving the performance for the target devices. The smaller the difference between the worst case and target device (smaller device variability), the smaller the device improvement for halo well architecture.

Circuit and Device Interactions

The choice of power supply (V_{CC}) and threshold voltage (V_T) will be critical in determining whether the performance of 0.1µm transistors can continue to be scaled. These parameters strongly affect chip active power, chip standby power, and transistor performance.

In this section, we review the power supply and threshold voltage scaling trends. We show that the loss in gate over drive (V_{CC}-V_T) is becoming so severe that this trend cannot continue without substantial loss in device performance. One possible solution that has been proposed is the use of dual threshold voltage transistors. It will be shown, however, that this will only extend the scaling trend by one technology generation at most.

V_{CC} and V_T Scaling

Figure 36 shows power supply and threshold voltage trends for Intel's microprocessor process technologies. As seen, the power supply is decreasing much more rapidly than threshold voltage. This has severe implications for device performance. Transistor drive current and therefore circuit performance is proportional to gate over drive (V_{CC}-V_T) raised to the power **n** where **n** is between 1 and 2 ($(V_{CC}-V_T)^n$).

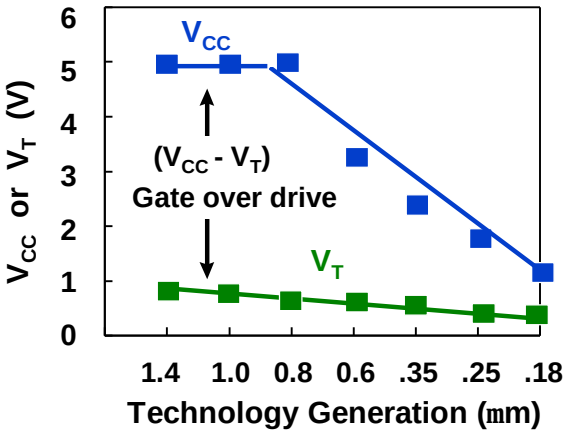

Figure 36: Power supply and threshold voltage scaling trend

In Figure 36, the gate over drive is shown to be rapidly decreasing for deep sub-micron devices, thereby strongly degrading device performance. As discussed previously, aggressive oxide, SDE, and well engineering are used to overcome the loss in gate drive and maintain the historical rate of transistor improvement.

To understand why these power supply and threshold voltages are being chosen, we need to understand chip active and standby power trends. Active power is set by circuit switching and is defined as $P = C_{LOAD} V_{CC}^2 f$ where **f** is the operating frequency and $\mathbf{C_{LOAD}}$ is the switching capacitance of the gate and wire load. Chip active power and frequency trends are shown for Intel's process technologies in Figure 37. Standby power results from junction and transistor sub-threshold source-to-drain leakage. For 0.1µm transistors, the sub-threshold leakage is the dominant contributor to standby power.

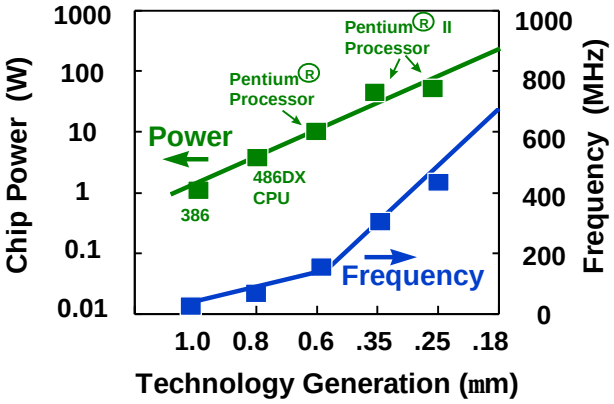

Figure 37: Chip power and frequency trends for Intel's process technologies

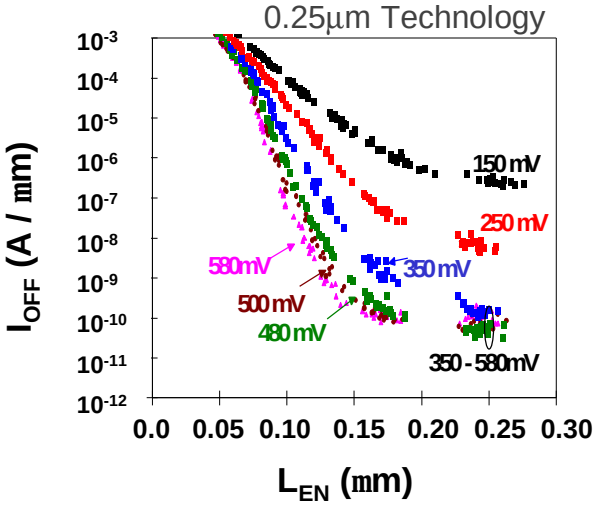

Figure 38: Off-state leakage versus channel length for 0.25µm transistors with different threshold voltage

Sub-threshold leakage is fundamental to silicon MOSFET operation and is set by the device threshold voltage. Sub-threshold off-state leakage versus channel length characteristics is shown in Figure 38. The active and standby power trends for Intel's process technologies are shown in Figure 39. In this figure, several interesting points can be observed. First, as microprocessor complexity increases, chip power is increasing to ~10-20W. Second, standby power for 1µm technology was .01% of active power, but is approaching 10% of active power in 0.1µm technologies. In order to limit the increase of standby power, threshold voltages need to increase. However, this increase strongly affects device performance because of reduced gate over drive. To maintain acceptable leakage values, the V_T's of transistors will need to increase by >0.25 V.

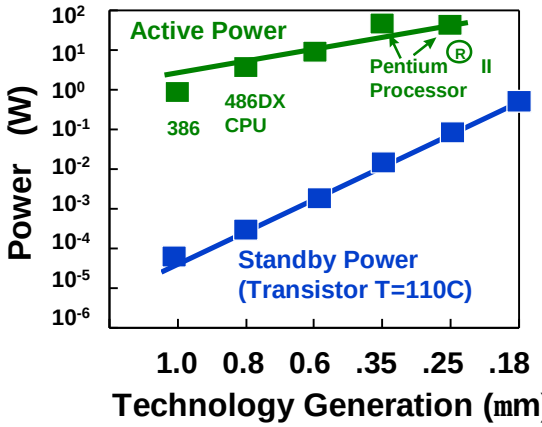

Figure 39: Active and standby power trends for Intel's technologies

Dual V_T Architecture

If power supply and threshold voltage scaling continues at the current trend, further reduction in gate overdrive will occur. A general rule for high performance transistor design is to maintain a V_{CC}/V_T ratio of at least four. A ratio of four provides a gate swing of one V_T to turn the device off and three V_T to drive the device. Figure 40 plots the V_{CC}/V_T ratio for Intel's previous technologies as well as the current projected trend. The projected scaling trend shows that beyond the 0.25µm technology, the ratio of V_{CC}/V_T will drop below 4. One technique to improve the gate drive and standby power trend is to offer circuit designers dual threshold voltage devices. This would consist of designing a high-performance, high-leakage, low-threshold voltage device and a low-performance, low-leakage, high-threshold voltage device. A chip would be designed such that only the critical paths would use the high-performance/high-leakage devices.

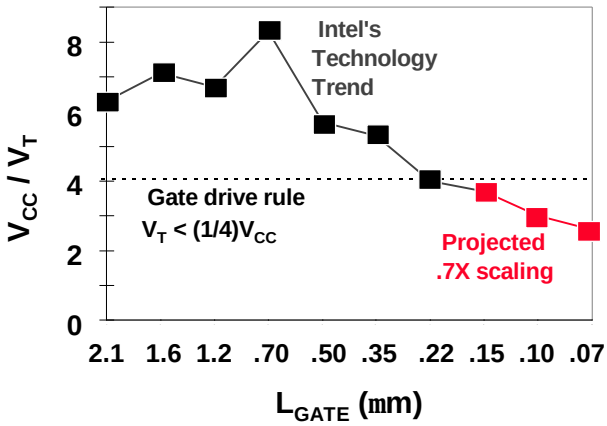

Figure 40: V_{CC}/V_T trend for Intel's process technologies

Figure 41 shows the performance and leakage current tradeoff for 0.25µm technology, lower threshold voltage devices. A 100x increase in leakage current would be required to extend the present performance trend by one generation. Whether or not a 100x increase in leakage could be tolerated would depend heavily on the circuit architecture and power constraints of the chip.

Alternate Device Options

Many designers have proposed new device architectures to improve device and circuit performance. In this section, we evaluate three of the most widely explored options and discuss the potential advantages and disadvantages of each.

SOI Device

One technique proposed to improve CMOS performance is to fabricate the devices on a silicon on insulator (SOI) substrate. SOI devices are classified into two types depending on the extent of the channel depletion layer (partially depleted or fully depleted) compared to the silicon thickness (T_{Si}). Fully depleted devices are not practical for deep sub-micron devices since the silicon thickness needs to be ~10.0 nm to control short channel effects. This silicon thickness is extremely difficult to manufacture and causes large device external resistance due to shallow SDE depths. Partially depleted devices are more suitable for deep sub-micron devices. However, since the channel region of the silicon layer is not entirely channel depleted, a partially depleted device offers no advantage for short channel effects or channel length scaling.

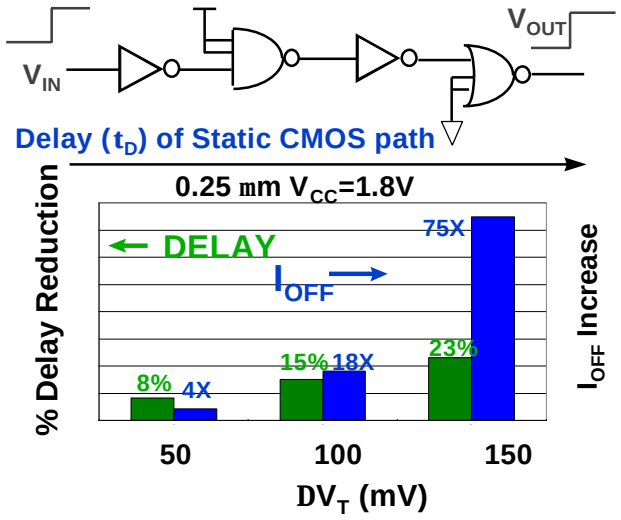

Figure 41: Performance and leakage current tradeoff for lower threshold voltage devices

Actually the partially depleted floating body can degrade short channel effects because of an uncontrolled lowering of V_T that is caused by impact ionization [28]. If the floating body can be controlled, partially depleted devices offer improvements in junction area capacitance, device body effect, and a gate-to-body coupling, which potentially results in a slightly larger drive current during switching.

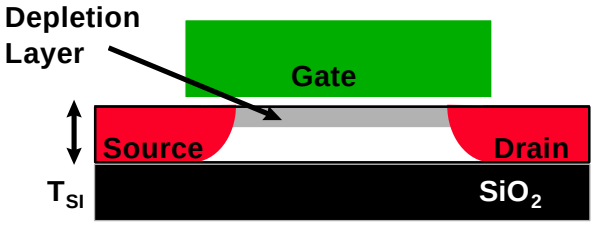

Figure 42: Cross section of an SOI device

Parameter	Best Case Gain
Junction Capacitance	12%
Body Factor	3%
Gate-to-Body Coupling	3%
Channel Length	0%
Total	18%

Table 3: Estimated improvement in circuit speed by device feature for a SOI device with unconstrained I_{OFF}

The best case estimated impact of these parameters on current generation circuit's speed improvements is shown in Table 3. We call it best case, since to date, no literature paper has demonstrated these device parasitic improvements without increasing the transistor off-state leakage. Studies done at Intel indicate that NMOS SOI devices require a somewhat higher threshold voltage than bulk devices to maintain an equivalent off-state leakage due to the floating body effect[28]. This higher threshold voltage offsets some of the other potential performance advantages of SOI. Also, in future high performance microprocessors where interconnect capacitances are becoming more dominant, the junction capacitance advantage of SOI will become less important. In summary, the performance gain going to the SOI architecture is less than one generation and will pose serious complications for circuit design due to floating body effects.

$Si_{1-x}Ge_x$ Channel Device

Another technique to improve transistor performance is to fabricate the device in a $Si_{1-x}Ge_x$ channel (see Figure 43). The $Si_{1-x}Ge_x$ channel region has been shown to increase hole mobility [29]. There are two reasons for the mobility gain: $Si_{1-x}Ge_x$ under compressive strain has improved mobility over Si; and the valence band offset between Si and $Si_{1-x}Ge_x$ localizes the hole inversion charge away from the SiO_2/Si interface, which reduces the effects of surface roughness scattering. Unfortunately, improving mobility becomes less important as the transistor is scaled into the deep sub-micron regime. This is due to the high lateral electric fields that cause the carrier velocity to saturate.

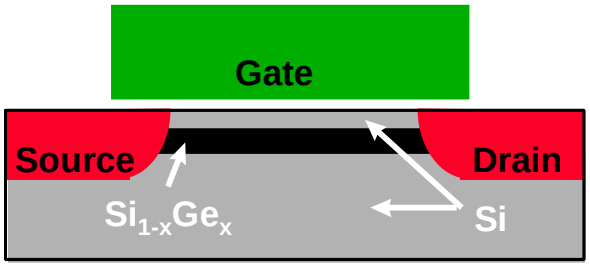

Figure 43: Cross section of a transistor fabricated with a $Si_{1-x}Ge_x$ channel

In Figure 44, the ratio of saturated drive current to mobility change is plotted for different device sizes. For long channel device lengths, the improvement in drive current is equal to the improvement in mobility. However, for deep sub-micron devices with channel lengths of ~0.1µm, a 4% improvement in mobility improves drive current by only 1%. If a $Si_{1-x}Ge_x$ channel improved electron or hole saturation velocity, there would be an improvement in drive current.

Unfortunately, electron and hole saturation velocities are similar if not slightly lower in SiGe than they are in silicon.

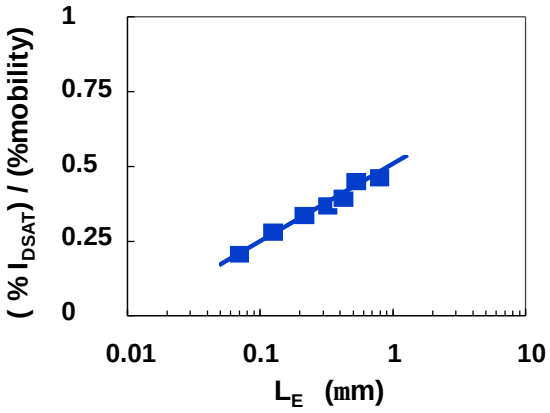

Figure 44: Ratio of I_{DSAT} change to mobility change versus channel length (for smaller devices, high electric fields cause velocity saturation)

Dynamic V_T Device

For low supply voltage operation (<0.6 V), a dynamic threshold voltage MOS device (DTMOS) has been proposed [30,31]. A DTMOS is formed by connecting the gate to the well as shown in Figure 45. This connection causes the threshold voltage of the device to be lowered during switching thereby increasing the transistor drive current. This technique is limited to supply voltages less than 0.6V to prevent the forward bias well-to-source junction from conducting large forward bias diode currents. The DTMOS technique has been proposed for devices fabricated on either bulk silicon or SOI substrates. Fabrication of these devices on SOI substrates is easier due to the electrical isolation of both n- and p-wells.

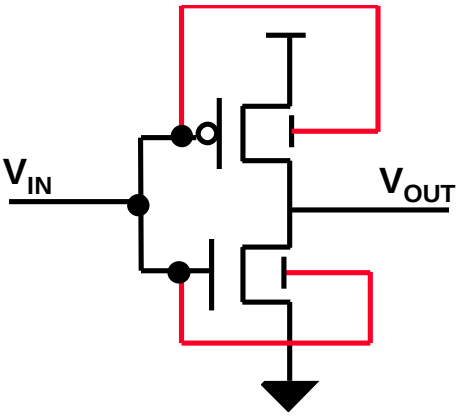

Figure 45: Circuit schematic of a dynamic threshold voltage MOS inverter

This technique can increase transistor drive current by over 20% through improved gate over drive (V_G-V_T). However, this technique offers little to no net gain over high performance, optimized, static V_T CMOS when differences in chip area are considered. When DTMOS is implemented on bulk silicon substrate (see Figure 46), there is a large performance degradation due to the increase in the switching load capacitance that is comprised of junction (C_J) and depletion (C_D) capacitance.

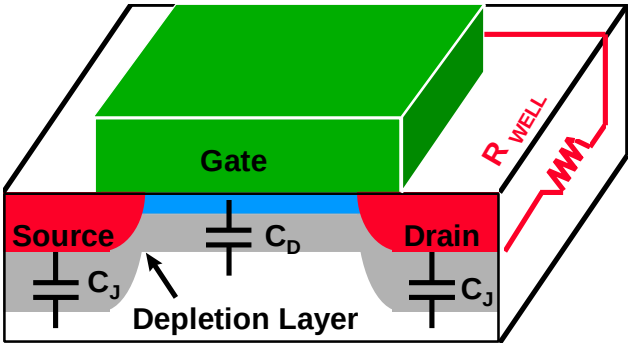

Figure 46: Transistor cross schematic of a dynamic threshold voltage MOS inverter

The performance degradation from the junction and depletion capacitance can be significantly reduced for DTMOS fabricated on an SOI substrate. However, for DTMOS on SOI, the RC time constant associated with the well resistance (R_{WELL}) and depletion capacitance (C_D) is not compatible with high frequency microprocessor applications. The R_{WELL}*C_D time constant can be ~1ns, which would consume half of the clock period for today's 500 MHz microprocessors. To minimize the RC delay associated with the poly-Si gate, companies have added metals to reduce the resistance to 2-3Ω/sq. By comparison, a DTMOS device in SOI can easily have a resistance component (R_{WELL}) on the order of 10^4-10^5 Ω/sq. or greater.

Although each of these alternate device structures has certain advantages, the overall device improvement is relatively small. In addition, manufacturing costs and circuit issues make it extremely difficult to justify the adoption of any of these device architectures.

Conclusions

Current performance scaling trends will not continue past the 0.13 - 0.10µm device technologies by using traditional scaling methods. Fundamental limits in SiO_2 scaling due to tunneling currents, in SDE junction depths due to large increases in external resistance, and in well engineering due to leakage constraints are currently being reached. At present, there is no clear alternate device architecture that has shown the potential for continuing the performance trends seen in the last 20 years. Aggressive exploration of high dielectric constant materials as well as developing a way to decrease SDE resistance offer the best hope for device and circuit improvements into the next century. These should be strongly supported.

Acknowledgments

The authors would like to acknowledge the collaborative efforts of our colleagues in the Portland Technology Development and Technology Computer Aided Design groups: T. Ghani, R. Rios, M. Stettler, M. Alavi, I. Post, S. Tyagi, R. Chau, M. Taylor, R. Nagisetty, J. Sandford, S. Ahmed, and S. Yang. The management support from R. Gasser, J. Garcia, S. Yang, L. Yau, and Y. El-Mansy is greatly appreciated.

References

[1] M. Bohr, S.U. Ahmed, L. Brigham, R. Chau, R. Gasser, R. Green, W. Hargrove, E. Lee, R. Natter, S. Thompson, K. Weldon and S. Yang, *IEDM Technical Digest*, 1994, p. 273.

[2] M. Bohr, S.S. Ahmed, S.U. Ahmed, M. Bost, T. Ghani, J. Greason, R. Hainsey, C. Jan, P. Packan, S. Sivakumar, S. Thompson, J. Tsai, and S. Yang, *IEDM Technical Digest*, 1996, p. 847.

[3] H.S. Momose, M. Ono, T. Yoshitomi, T. Ohguro, S-I. Nakamura, M. Saito, and H. Iwai, *IEDM Technical Digest*, 1994, p. 593.

[4] S.A. Hareland, S. Krishnamurthy, S. Jallepalli, C.-F. Yeap, K. Hasnat, A.F. Tasch, and C.M. Maziar, *IEDM Technical Digest*, 1995, p. 933.

[5] S.-H.Lo, D.A. Buchanan, Y. Taur, and W. Wang, *IEEE Electron Device Letter*, 1997, p. 209.

[6] S. A. Campbell, D.C. Gilmer, X.-C. Wang, M.-T. Hsieh, H.-S. Kim, W.L. Gladfelter, and J.Yan, *IEEE Electron Device*, 1997, p. 104.

[7] S. Kamiyama and T. Saeki, *IEDM Technical Digest*, 1991, p. 827.

[8] C.G. Parker, G. Lucovsky, and J.R. Hauser, to be published, 1997.

[9] H.-H. Tseng, P.G.Y. Tsui, P.J. Tobin, J. Mogab, M. Khare, X.W. Wang, T.P. Ma, R. Hegde, C.Hobbs, J.Veteran, M. Hartig, G. Kenig, V. Wang, R. Blumenthal, R. Cotton, V. Kaushik, T. Tamagawa, B.L. Halpern, G. J. Cui, and J. J. Schmitt, *IEDM Technical Digest*, 1997, p. 647.

[10] S. Thompson, *VLSI Symposium Technology Short Course*, 1998.

[11] M. Ono, M. Saito, T. Yoshitomi, C. Fiegna, T. Ohguro, and H. Iwai, *IEDM Technical Digest*, 1993, p. 119.

[12] A. Hori, H. Nakaoka, H. Umimoto, K. Yamashita, M. Takase, N. Shimizu, B. Mizuno and S. Odanaka, *IEDM Technical Digest*, 1994, p. 485.

[13] P.A. Packan and J.D. Plummer, *Applied Physics Letter*, 1990, p. 1787.

[14] D.F. Downey, C.M. Osburn, S.D. Marcus, *Solid State Technology*, 1997, p. 71.

[15] D.J. Eaglesham, P.A. Stolk, H.-J. Gossmann, and J.M. Poate, *Applied Physics Letter*, 1994, p. 2305.

[16] A.D. Lilak, S.K. Earles, K.S. Jones, M.E. Law, *IEDM Technical Digest*, 1997, p. 493.

[17] M.E. Law and K.S. Jones, *Proceedings of the Process Physics Symposium of the Electrochemical Society*, 1996, p. 374.

[18] S.E. Thompson, P.A. Packan, and M.T. Bohr, *VLSI Symposium Digest*, 1996, p. 154.

[19] S. Venkatesan, J.W. Lutze, C. Lage and W.J. Taylor, *IEDM Technical Digest*, 1995, p. 419.

[20] M. Rodder, S. Aur, And I.-C. Chen, *IEDM Technical Digest*, 1995, p. 415.

[21] J.B. Jacobs and D. Antoniadis, *IEEE Transactions Electron Devices*, 1995, p. 870.

[22] G.G. Shahidi, J.D. Warnock, J. Comfort, S. Fischer, P.A. McFarland, A. Acovic, T.I. Chappell, B.A. Chappell, T.H. Ning, C.J. Anderson, R.H. Dennard, J.Y.C. Sun, M.R. Polcari, and B. Davari, *IBM Journal Research Development*, 1995, p. 229.

[23] M. Cao, P. Griffin, P. Vande Voorde, C. Diaz, and W. Greene, *VLSI Symposium Digest*, 1997, p. 85.

[24] C.T. Sah, *Fundamentals of Solid-State Electronics*, 1991, p. 553.

[25] C.F. Codella and S. Ogura, *IEDM Technical Digest*, 1985, p. 230.

[26] T. Hori, *IEDM Technical Digest*, 1994, p. 75.

[27] Y. Taur and E.J. Nowak, *IEDM Technical Digest*, 1997, p. 215.

[28] R. Chau, R. Arghavani, M. Alavi, D. Douglas, J. Greason, R. Green, S. Tyagi, J. Xu, P. Packan, S. Yu, and C. Liang, *IEDM Technical Digest*, 1997, p. 591.

[29] K. Ismail, J.O. Chu, and B. S. Meyerson, *Applied Physics Letter, vol. 64*, 1994, p. 3124.

[30] F. Assaderaghi, D. Sinitsky, S. Parke, J. Bokor, P.K. Ko, and C. Hu, *IEDM Technical Digest*, 1994, p. 809.

[31] A. Shibata, T. Matsuoka, S. Kakimoto, H. Kotaki, M. Nakano, K. Adachi, K. Ohta, and N. Hashizume, *VLSI Symposium Digest*, 1998, p. 76.

Authors' Biographies

Scott Thompson joined Intel in 1992 after completing his Ph.D. under Professor C. T. Sah at the University of Florida on thin gate oxides. He has worked on transistor design and front-end process integration on Intel's 0.35, 0.25, and 0.18µm silicon process technology design for the Pentium® and the Pentium® II microprocessors. Scott is currently managing the development of Intel's 0.13µm transistor design. His email address is scott.thompson@intel.com.

Paul Packan received his Ph.D. degree in Electrical Engineering in 1991 from Stanford University. He joined Siemens AG in Munich Germany in 1991 working in the area of high speed bipolar transistor architecture. In 1992 he joined Intel Corp. working in the field of process and device simulation for MOS devices. He worked on the development of the 0.35, 0.25 and 0.18 um technologies and is currently managing the process and device modeling group. His email address is paul.a.packan@intel.com.

Mark T. Bohr joined Intel in 1978 after receiving a MSEE from the University of Illinois. He has been a member of the Portland Technology Development group since 1978 and has been responsible for process integration and device design on a variety of DRAM, SRAM, and logic technologies, including recently 0.35µm and 0.25µm logic technologies. He is an Intel Fellow and director of process architecture and integration. He is currently directing development activities on 0.18µm and 0.13µm logic technologies. His email address is mark.bohr@intel.com.