Scaling Of Si MOSFETs For Digital Applications

By Dustin K. Slisher, Ronald G. Filippi, Jr., Daniel W. Storaska and Alberto H. Gay

Final Project in the "Advanced Concepts in Electronic and Optoelectronic Devices"
class of Professor M. S. Shur
12/10/99

TABLE OF CONTENTS

Page

I. ABSTRACT

As the demand grows for high performance and high density integrated circuits, MOSFET scaling to submicron regimes will continue to be at the core of device and circuit design. While MOSFET dimensions are reduced, circuit requirements demand maintaining long channel behavior and minimizing short channel as well as parasitic effects. At the same time, higher current driveability requires thinner gate oxides in shorter channel length devices. In this project, we give a detailed literature review on the subject of MOSFET scaling. We discuss various scaling approaches, the effect of scaling on initial device characteristics, the limits imposed by reliability concerns in scaled-down MOSFET technologies, techniques to control short channel effects and unconventional approaches to MOSFET scaling.

II. INTRODUCTION

Silicon (Si) based integrated circuits (ICs) have become the backbone of today' s semiconductor world. Technology has progressed much since the days of vacuum tubes. The semiconductor industry went through a series of increasingly smaller device sizes commonly known today as SSI, MSI, LSI, ELSI, VLSI. (Here, SI stands for scale integration and the first letter stands for a size from small to very large.) The VLSI circuit first appeared in 1981, and since then, the industry has grown so much that it "is now the largest industry in terms of output as well as employment in many nations," [1]. Integration, that is, the number of transistors per chip, has increased over five orders of magnitude through the previously mentioned generations, while computation capability has increased by at least three [2]. The microelectronics industry has a huge impact in the world in terms of economic, social, and political development.

Microelectronics has grown in large part because of its ability to continually improve performance while reducing costs. There is a constant drive to make devices that occupy less space, consume less power and have shorter delays. During the past thirty years the minimum feature size has improved by close to two orders of magnitude. These small features are driven by several main goals. Smaller features mean larger device density, which in turn equates to less raw material for the same amount of processing power. This results in lower cost per MIPS (Million instructions per second), diminished transit times, shorter time delays and improved performance. As an example of improved performance over short periods of time, PCs 5 years ago were running at 33 Mhz based on 386 or 486 processors. Today, cutting edge Pentium IIs are running at 450 Mhz (>10x improvement). Having smaller dimensions also means lower power consumption per device. The challenges related to smaller device dimensions are numerous and shall be discussed later in detail

The heart of the Si integrated circuit is the transistor. There are two main transistor technologies in the market today: Bipolar and CMOS (Complementary Metal Oxide Semiconductor). For a period of time, Bipolar technology offered better performance, but consumed at least an order of magnitude more power than CMOS at comparable performance. Increased power consumption not only drives increased power costs, but higher cost and complexity in cooling hardware. This includes extended surface fins, fans, water cooling, or even liquid nitrogen. An additional disadvantage of this is the floor space consumed by the cooling equipment. It is clear then, for all IC technologies the need for decreased power consumption. Today, CMOS is the dominant IC technology due largely to comparable performance and improved power efficiency [3,4].

CMOS technology makes use of both n- and p-channel MOSFETs (Metal Oxide Semiconductor Field Effect Transistors), as illustrated in Figure 1. Fabrication of a chip begins with a single Si crystal wafer. Impurities, such as boron (called an acceptor because of its ability to accept electrons) or phosphorus (called a donor), are introduced into the Si matrix to create hole-dominated (p) or electron-dominated (n) regions, respectively. Oxide is thermally grown on top of the channel between the self-aligned drain and source n^+ regions (for n-MOSFETs) and p^+ regions (for p-MOSFETs). A heavily-doped polysilicon gate is formed on top of the oxide. By applying an appropriate voltage to the gate, the current between the source and the drain regions is modulated accordingly.

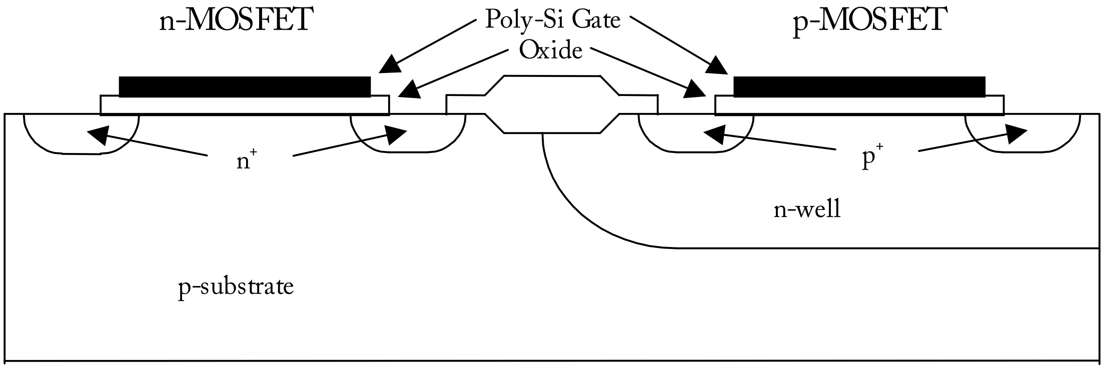


Figure 1. Schematic cross section of n- and p-channel MOSFETs used in CMOS n-well technology.

Today, the leading edge CMOS technology has a minimum channel length under 0.15 μm and a gate oxide thickness less than 35 Å . Roldan, et al. [5], in October 1998 considered analytically and experimentally a MOSFET with a channel length of only 0.07 μm. Intel has also experimented with a 0.06 μm gate length transistor [4]. We are entering the age of ULSI (ultra large scale integration), that is, scaling in the nanometer range. Scaling is the process by which device dimensions are made smaller and will be explained in more detail below.

In 1994 the Semiconductor Industry Association introduced the technology roadmap [1,6]. It was based on the assumption that the industry would continue its advancement at an historical pace of a new generation every three years-

Table I. MOSFET scaling trend for high performance – 1994 Roadmap [1,6].

Year	1991	1994	1997	2001	2005	2009
Minimum feature size (μm)	0.5	0.35	0.25	0.18	0.13	0.09
SRAM density	4M	16M	64M	256M	1G	4G
V_{CC} (V)	5	3.3	3.3	3.3	2.2	2.2
Gate oxide thickness (nm)	13.5	9	8	7	4.5	4
Junction depth (μm)	0.15	0.15	0.1	0.08	0.08	0.07
Effective channel length (μm)	0.40	0.30	0.25	0.2	0.15	0.13
Threshold voltage (V)	0.7	0.7	0.7	0.7	0.7	0.7
NMOS I_{DSAT}@ V_{gs}= V_{cc} (mA/μm)	0.64	0.48	0.55	0.65	0.51	0.57
PMOS I_{DSAT}@ V_{gs}= V_{cc} (mA/μm)	0.31	0.22	0.26	0.32	0.24	0.28

Another piece of the technology roadmap addresses these same factors for low power consumption. Our emphasis in this paper will be the high performance side since improved performance usually outweighs low power consumption. We are ahead of this roadmap since Intel has already introduced microprocessors based on 0.18 μm device technology. The pace has been increased to a new generation every two years. Therefore, in 1997, a new technology roadmap [7] incorporating these changes was introduced (see Table II) and a 1999 version is already on the drawing board.

Table II. MOSFET scaling trend for high performance – 1997 Roadmap [7].

Year	1997	1999	2001	2003	2006	2009	2012
Minimum feature size (μm)	0.25	0.18	0.15	0.13	0.10	0.07	0.05
SRAM density	64M	256M	1G	1G	4G	16G	64G
V_{CC} (V)	1.8-2.5	1.5-1.8	1.2-1.5	1.2-1.5	0.9-1.2	0.6-0.9	0.5-0.6
Gate oxide thickness (nm)	4-5	3-4	2-3	2-3	1.5-2	<1.5	<1.0
Junction depth (nm)	50-100	36-72	30-60	26-52	20-40	15-30	10-20
NMOS I_{DSAT}@ V_{gs}= V_{cc} (mA/μm)	0.60	0.60	0.60	0.60	0.60	0.60	0.60
PMOS I_{DSAT}@ V_{gs}= V_{cc} (mA/μm)	0.28	0.28	0.28	0.28	0.28	0.28	0.28

There are some serious concerns in terms of physical factors limiting the scaling capabilities [8]. These factors will be discussed in more detail below, but at this time it is appropriate to briefly visit them. The physical concerns of scaling include, but are not limited to, increased leakage currents, limits to doping and decreased mobilities with higher doping, limits to minimum allowable oxide thickness imposed by direct tunneling, effects of contact resistance, constant energy gap, difficult scalability of threshold voltages, DIBL, GIDL, power dissipation, etc. Reliability concerns include hot carrier degradation, gate dielectric breakdown, and interconnect reliability. Fabrication concerns include lithography and contamination. Some of these limits are shown in the following table from [4]:

Table III. Scaling limits for MOSFET technologies.

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	0.06 μm	Leakage (I_{off})
Gate length	0.10 μm	Leakage (I_{off})

There has been a lot of research into different forms of scaling such as constant field, constant voltage, electrostatic, subthreshold and off-current scaling. There is also a consistent drive to find the real limitations of Si. Some of the Si research that seems promising includes dual gate devices, in which there are two depletion regions on the channel to give added current control. Conversely, there are groups working on alternative technologies, such as silicon germanium or other semiconductor materials, which provide comparable performance at lower bias and power consumption. Silicon on insulator (SOI) also demands attention since it offers characteristics that may solve shallow junction, soft error, and isolation problems [3]. Low temperature CMOS is still on the table despite the obvious cost and complexity drawbacks of cryogenic cooling.

There is no telling what the future holds. The current estimates put the maximizing of CMOS technology around 2010, yet the industry has faced almost impossible odds in the past and succeeded. It is clear that any material will pose limitations, and Si is no exception. But there are so many materials which are not fully understood today. It is quite apparent that the physical limits are far from being attained.

III. LITERATURE REVIEW

A. Scaling Approaches

Scaling is the process of miniaturizing devices while attempting to maintain electrical characteristics constant. There have been many different attempts at scaling. The main problem with miniaturization is the direct, and more importantly, indirect dependence of electrical characteristics on controllable physical parameters. This causes many non-ideal effects that hinder the performance or power consumption characteristics of devices.

The first complete scaling scheme was introduced by Dennard, et al., in 1974 [9]. The method is called constant electric field scaling (see Table IV). In order to scale down the depletion region, internal fields, currents, and capacitances, among others, all dimensions are scaled by a factor K. Depending on the variable, the parameter could be multiplied, or divided by K. In doing so, these non-ideal effects were avoided to a certain extent. The main drawback of this scaling scheme is that it is often not possible to scale parameters in the required proportions. For example, substrate doping has an upper limit of 10^{18} cm^{-3}. So, if the limit is already reached, further doping is impossible. Threshold voltage scaling poses some particularly challenging problems. Working devices of 0.25 μm channel length or longer have roughly a 0.7 V threshold voltage, while experimental devices of 0.1 μm channel length have 0.33 to 0.40 V threshold voltage. Constant field scaling is clearly only approximated, not followed exactly.

Different scaling schemes were soon to follow, such as constant voltage scaling (see Table IV). Constant voltage scaling attempts to address limitations imposed by industry convention. In constant electric field scaling, the source voltages are decreased by a factor of K. As shown in the technology roadmaps (see Tables I and II), the industry has agreed, years in advance, on what the supply voltages will be, thus providing manufacturers enough lead time to design and manufacture power supplies. Designing and manufacturing unique power supplies for each particular application or channel length is not practical or economical since it requires too much time and money for the resulting performance improvement. Therefore, it becomes inevitable to accept standard power supply voltages when designing a device. Constant voltage scaling is therefore a more practical application of the more ideal method of constant electric field scaling

[10]. One drawback of this method is that by not scaling the supply voltage higher fields are created in the device. This leads to mobility degradation, hot carrier effects, and other reliability problems. Also, this method consumes more power and requires better cooling methods than constant electric field scaling.

A third proposed method is constant electrostatic scaling, or quasi-constant voltage (see Table IV). In this method, dimensions are scaled by the same factor K, but potentials are scaled by a different factor $\lambda = K^{0.5}$ [11]. This method is another compromise between reality and ideal constant electric field scaling. The factor λ is applied when the voltages cannot be reduced by K. This leaves the field pattern constant and reduces the effects of punch through and DIBL. While this method addresses most of the practical challenges of the previous two, it remains a theoretical method and serves only as a good starting point for device designers. Further testing and optimizing for a particular application will always be required.

Table IV. Scaling Laws at 300 K.

Parameter	Constant Field Scaling	Constant Voltage Scaling	Constant Electrostatic Scaling
Gate length	1/K	1/K	1/K
Gate width	1/K	1/K	1/K
Gate oxide	1/K	1/K	1/K
Junction depth	1/K	1/K	1/K
Doping density	K	K^2	K^2/λ
Drain voltage	1/K	1	$1/\lambda$
Drain current	1/K	K	K/λ^2
Threshold voltage	1/K	1	$1/\lambda$
Propagation delay time	1/K	1/K	1/K
Supply voltage	1/K	1	$1/\lambda$
Gate capacitance	1/K	1/K	1/K
Line current density	K	K^3	K^3/λ^2
Number of transistors	K^2	K^2	K^2
Chip size	1	1	1
Power density	1	K^3	K^3/λ^3

A fourth scaling method was proposed by Brews [12,13]. This method is called subthreshold scaling and is empirical. While it does not dictate specific factors for scaling individual dimensions, it provides a framework for which combinations of parameters will result in long or short channel behavior. The design criterion selected to represent long channel behavior in a scaled down device is defined as having a variation of less than 10% in the drain current per 0.5 V variation in drain-source voltage. The freedom this method brings is that it does not start with a large working device of fixed dimensions, but rather it allows independent manipulation of a large number of variables as long as the remaining variables compensate for these changes. The process is defined as follows in Figure 2.

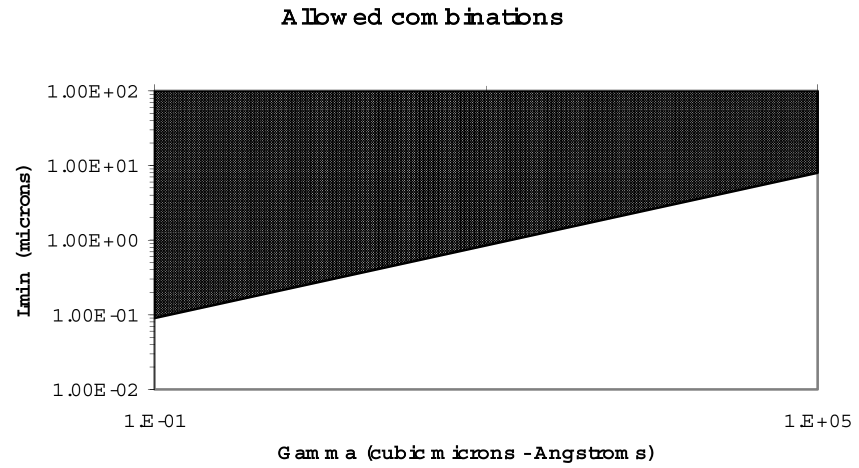


Figure 2. Log-log plot of L_{min} versus Gamma, γ, for subthreshold scaling .

The line in Figure 2 is described by:

$$L_{MIN} = 0.4\gamma^{1/3} \qquad \textbf{Equation (1)}$$

$$Gamma = \gamma = r_j d(W_S + W_D)^2 \qquad \textbf{Equation (2)}$$

where r_j is the junction depth in microns (µm), d is the gate oxide thickness in angstroms (Å), W_S is the width of the source depletion region in microns (µm) and W_D is the width of drain depletion region in microns (µm). Combinations above the line in Figure 2 (shaded region) exhibit long channel subthreshold behavior, while those under the line exhibit short channel subthreshold behavior. This method has additional drawbacks in that it is not understood why it works and it has not been tested for gate lengths < 0.3 µm [12]

The final scaling method, also by Brews [13], is called Off-current scaling and is more complex in practice than the previous methods. In this method the doping profile characteristics are varied in order to obtain an acceptable combination of off current, I_{off}, and threshold voltage for a minimum channel length device [10,13]. First, the scaling assumes that source voltage, gate material, oxide thickness, and junction depth have been previously fixed by system or processing constraints. The minimum length is fixed by the lithographic process. It is evident that this method, while a good check for a design that is partially complete, lacks flexibility of the aforementioned variables in successive iterations and cannot be considered as a stand alone scaling strategy. The doping, centroid of the threshold voltage ion implant (x_c) and exposed dose (D_I) are determined assuming the channel behaves as a long channel for an initial approximation. Since we know this is not the case, the acceptable I_{off} is deliberately picked smaller than the needed value in the device being designed to compensate for an I_{off} increase due to DIBL. DIBL current is determined by any particular model the designer chooses to employ. Then, a minimum substrate doping is determined by setting a very low limit on subthreshold punch through at a selected maximum drain-source voltage. Finally, D_I and x_c are reviewed to keep the channel depletion region as small as possible, that is, to keep the shift in the threshold voltage due to short channel effects as small as possible for a given minimum length. Therefore, this process does not attempt to eliminate the short channel effects like previous methods, but rather to design a device that compensates for effects of DIBL, punch through, etc. While the effects are still noticeable (the application does not require total absence of them), they do not affect the designed performance of the device since they were taken into consideration in the design. Threshold voltage and long channel off current tradeoffs are determined by using a series of curves with fixed doping levels and oxide thicknesses. The y-axis is the threshold voltage and the x-axis is the ratio of free carrier density with zero gate-source voltage to the free carrier concentration at the midgap, that is the point of field induced polarity reversal. These graphs must be built for each different oxide thickness and substrate doping and are of limited utility and shall not be included in this paper.

We can see that all scaling methods are attempts to replicate long channel behavior in a short channel device. No scaling method provides an exact solution, and designing a device requires many iterations, experience and perhaps, artistic ability, on the part of the designer. The best approach may very well be combining one of the first three methods with one or more of the latter two. Many of these methods are compromises between reality and ideal (constant electric field) scaling. All of the methods attempt to keep proportions between physical and electrical characteristics of the devices constant, thereby avoiding short channel and non-ideal effects. There clearly is a lot of room for new and improved scaling methods to emerge which will hopefully address the flaws of the current available alternatives without increasing the complexity of the method to unmanageable limits. Still, there has been, and will continue to be significant, sometimes surprising, progress in scaling devices; the limits of current theory have not been reached. Perhaps, in the future we will need different theories.

B. Effects of Scaling on Initial Device Characteristics

1. Introduction to Scaling Effects

Since the integrated circuit era began in 1959 the gate length has been decreasing [14]. While the gate length will continue to decrease for some time, the shortest gate length is still an unknown. There are two main driving reasons to decrease the minimum feature size of the MOSFET, density and speed [10]. The definition of a short channel is when the gate length is on the same magnitude as the depletion region of the drain and source junctions. Also the short channel MOSFET can be defined when the effective channel length, L_{eff}, is approximately the same length as the source and drain implant depth x_j [15]. An empirical formula for short channel effects is given by equation (1), where L_{min} is the minimum length the gate can be before short channel effects have to be considered. As devices are decreased some effects will become dominant. For example, as the gate length is decreased channel-length modulation (i.e., the dependence of the effective channel length on the drain bias) will have more of an effect on channel current.

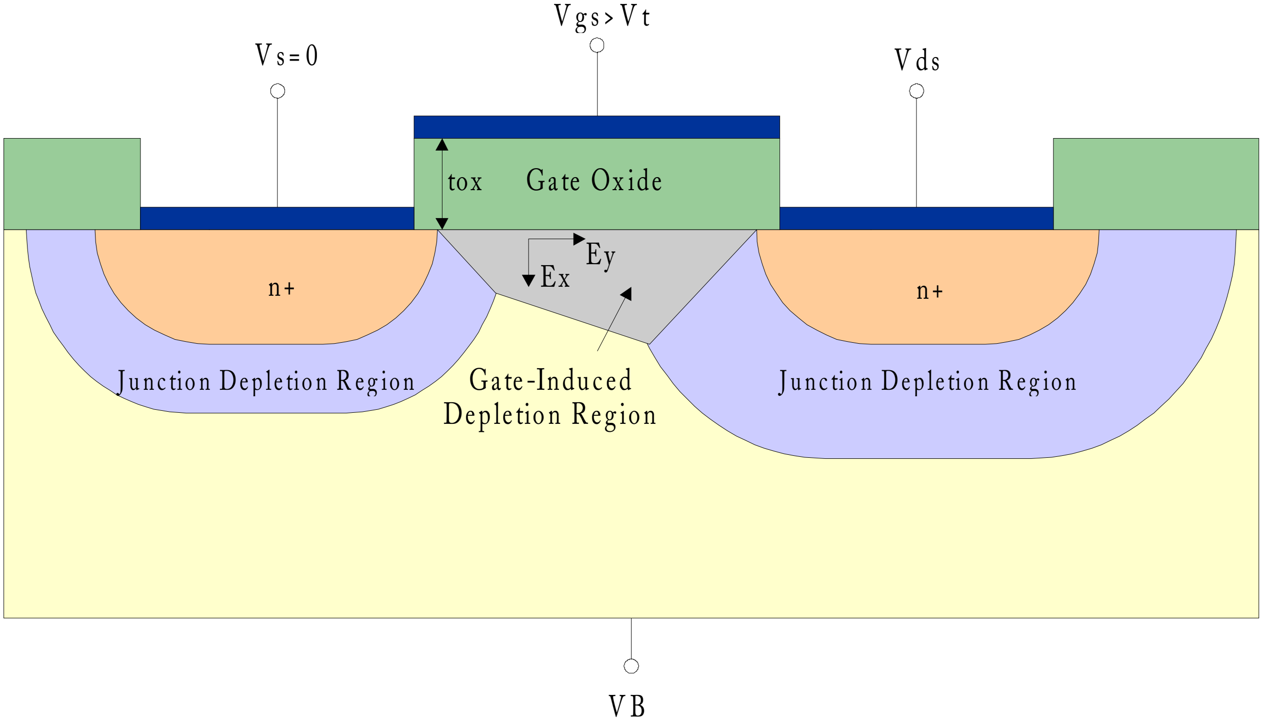


Figure 3. Short Channel MOSFET [15].

2. Velocity Saturation

The electric field E_y between the source and drain increases as the gate length decreases. The electron drift velocity is proportional to low electric fields perpendicular to the gate. As E_y is increased the electron velocity saturates. Therefore the current in the channel saturates. When E_y approaches 10^5 V/cm, the electron velocity saturates at 10^7 cm/s. This saturation can have great impact on the current-voltage properties. Consider the drain-source current, I_{ds}, of a MOSFET in saturation mode ($V_{ds} \geq V_{gs} - V_t$) [15]

$$I_{ds}(\mathrm{sat}) = W \cdot vd(\mathrm{sat}) \cdot \int_0^{Leff} q \cdot n(x)dx = W \cdot vd(\mathrm{sat}) \cdot |Q_I| \qquad \textbf{Equation (3)}$$

Since V_{ds}=V_{Dsat} (where V_{Dsat} is the voltage at which the velocity saturates) the current equation is as follows [15]:

$$I_{ds}(\mathrm{sat}) = W \cdot vd(\mathrm{sat}) \cdot C_{ox} \cdot V_{DSAT} \qquad \textbf{Equation (4)}$$

The I_{ds}(sat) current using this equation is lower than using the normal long channel equation for I_{ds}(sat). The saturation current is no longer a quadratic function of V_{gs} and is primarily independent of channel length [15]. The following table lists proposed V_{ds} values based on a saturation velocity for Si of approximately 1e5 V/cm

Table V.

L (μm)	Vds(V)
0.5	5
0.35	3.5
0.25	2.5
0.175	1.75
0.15	1.5
0.10	1.0

3. Threshold Voltage, Vt

a. Length Dependencies

The decrease in Vt is a clear indicator of short channel effects [16]. Plotting L_{eff} on the x-axis and threshold voltage on the y-axis makes a Vt roll-off chart. The chart indicates the minimum L_{eff} that will be acceptable. Vt roll-off is one of the most serious consequences of short channel effects [16]. Figure 4 below demonstrates Vt roll-off for n- and p-channel MOSFETs [17].

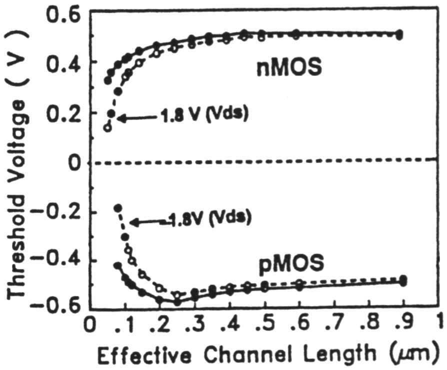


Figure 4. Vt versus L_{eff} demonstrating Vt roll-off [17].

The threshold voltage (Vt) of MOSFETs cannot continue to be scaled down as the gate length is decreased. The subthreshold I_{off} increases as the Vt is decreased. An increase in I_{off} is a serious threat to the continued performance enhancements of the MOSFET transistor. For transistors less than 0.25 μm designers must consider the trade off between speed and lower power consumption [18].

As the gate length is reduced below 2 μm, the long channel approximation for the threshold voltage is not as accurate. The Vt generally decreases as the gate length is decreased (this is known as Vt roll-off). Also, the Vt decreases as the drain-source voltage (V*ds)* is increased. In order to predict the Vt of a short channel device, the shift in the threshold voltage, ΔVt, must be approximated [10]. The short channel effect (SCE) Vt is given by the following formula where *Vto* is the long channel Vt [19]:

$$(Vto)_{SCE} = Vto - (\Delta Vt) \qquad \textbf{Equation (5)}$$

A method of calculating ΔVt is by using the charge-sharing model. This model assumes that the charge under the gate is shared between the source/drain depletion regions and gate inversion region. Therefore less voltage is required to invert the channel. ΔVt increases as the depletion region's length approaches the gate length. Yau in 1974 proposed a simple model to predict ΔVt. The model used the assumption that the charge caused from the gate is simply the trapezoidal region under the gate as in Figure 3. The following analytical formula to calculate ΔVt can be used for uniformly doped channels [10]:

$$\Delta Vt = \frac{q \cdot N_{SUB} \cdot d_{\max} \cdot r_j(\sqrt{\{1+\frac{2d_{\max}}{r_j}\}} -1)}{C_{ox} \cdot L} \qquad \textbf{Equation (6)}$$

d_{max} is the maximum width of the depletion region under the gate, and r_j is the length of the depletion region of the source/drain. The model shows that decreasing the gate oxide thickness ($C_{ox}=\varepsilon_{ox}/t_{ox}$) and decreasing the depletion regions of the source/drain will decrease ΔVt. The model works well to understand the concept of decreasing ΔVt, but does not predict the change in the Vt accurately against experiment data, especially for narrow gate lengths and high Vds. The model does provide a first order approximation [10].

DIBL (Drain Induced Barrier Lowering) was introduced in 1979 by Troutman [10]. As the voltage drop between the source and drain increases, the depletion region under the drain can lower the potential barrier from the source-to-channel junction. If the barrier between the source and channel is decreased electrons are more freely injected into the channel region. Therefore the threshold voltage is lowered and the gate has less control of the channel current [20].

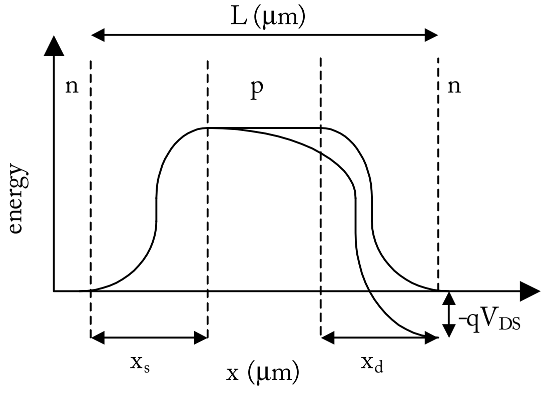


Figure 5. DIBL (Drain Induced Barrier Loweing).

Refer to Figure 5, which illustrates the DIBL effect for an n-channel MOSFET. The current in the channel depends exponentially on the barrier height. A slight decrease in the barrier height can have a significant impact on the channel current. A model has been developed to calculate the change in ΔVt caused from DIBL. The theoretical solution would require a two dimensional solution of Poisson's equation. The following equation was derived using a more simple analytical calculation [21]:

$$\Delta Vt = -\eta \Delta \psi(x_s) = -\eta \cdot V(x_s) = -\sigma V_{DS} \qquad \textbf{Equation (7)}$$

where

$$\sigma \approx \frac{2\eta\chi d^0_{dep}}{\lambda} \frac{\sinh\left(\frac{x_s}{\lambda}\right)}{\cosh\left(\frac{L - x_d}{\lambda}\right) - \cosh\left(\frac{x_s}{\lambda}\right)} \qquad \textbf{Equation (8)}$$

$$\lambda = d^0_{dep}\left(1 + \frac{\varepsilon i}{\varepsilon s} \cdot \frac{d^0_{dep}}{di}\right)^{-1/2} \qquad \textbf{Equation (9)}$$

d^0_{dep} is the depletion width when V=0, di is the insulator thickness, Ψ is the potential in the channel, εi and εs are the electrical permittivities of the insulator and silicon, respectively, and x_d and x_s are the depletion region depths for the drain and source, respectively.

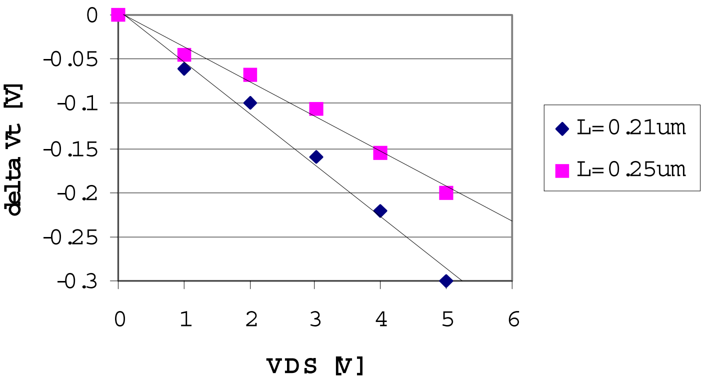


Figure 6. ΔVt versus VDS for two different channel lengths [21, 22].

Experimental data agrees well with equation (7), which indicates a linear relationship between ΔVt and $V_{DS.}$ Figure 6 shows a linear relationship for two channel lengths, where the slope of the lines is σ. The data was collected by Chung, et al [21,22].

A way to measure DIBL is by measuring the threshold voltage in the linear region with Vds= 0.05 volts and measuring the threshold in saturation with Vds between 1.2-3.0 volts depending on L design. As was mentioned above the threshold voltage will decrease with increasing Vds. To graphically show this, subtract saturated Vt from linear Vt to find delta Vt, and plot this difference verses L_{eff}.

b. Width Dependencies

Another effect on threshold voltage occurs as the width is scaled. The width scaling effect is not as severe as the length scaling effects. Three effects will be presented here as the width is scaled. Two cause the Vt to increase (opposite to length scaling) and the other causes the Vt to decrease. The first two effects are caused from fabrication of isolation structures, either raised field-oxide or semi-recessed LOCOS (Local Oxidation of Silicon). Raised field oxide is created by first growing the gate oxide and removing the oxide over the source and drain implant regions, as seen in Figure 7(a). Semi-recessed LOCOS is shown in Figure 7(b). The third effect is caused from fully-recessed-LOCOS [10].

The first effect considers the depletion region perpendicular to the current flow from source to drain along the gate edge in the L direction. The electric field from the gate causes depletion in the vertical direction and consequently in the lateral direction also. The depletion region parallel to the current flow in the source to drain direction will be discussed in the next section, where it will be shown to reduce the threshold voltage. But the other depletion region causes the threshold voltage to shift up. The bulk charge in the channel is actually higher when considering the charge from the lateral and vertical depletion regions. As the gate width is reduced the proportion of lateral depletion charge will become a larger percent than from a wider gate. If the lateral depletion charge remains constant regardless of gate length, it is believed that a higher gate voltage will be needed to invert the channel since the total depletion region will be effectively larger [10].

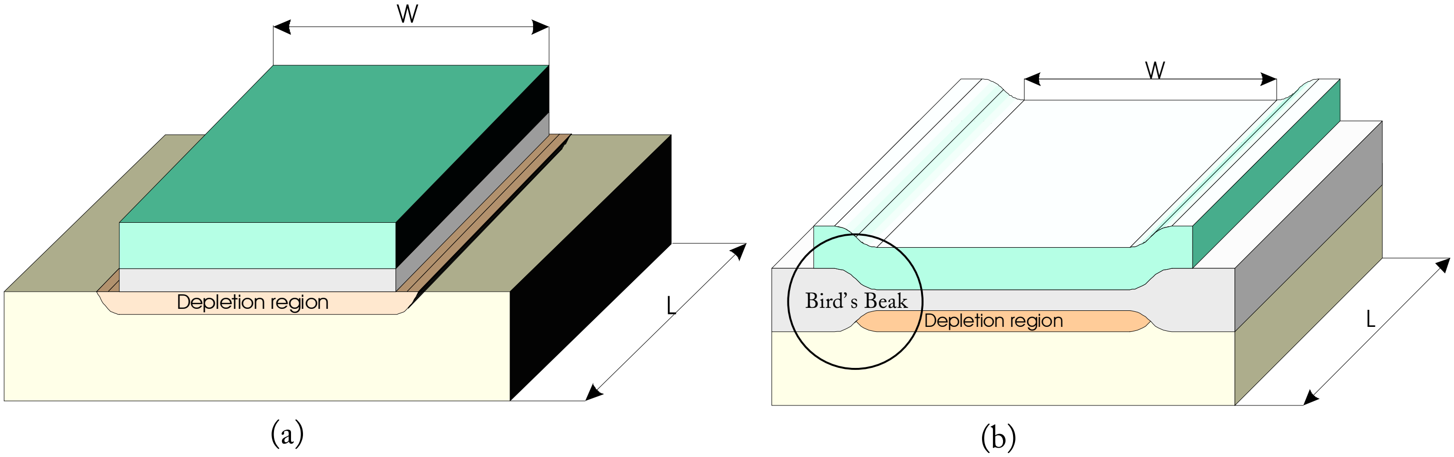


Figure 7. (a) Raised field oxide process and (b) semi-recessed LOCOS process [10].

The second effect is from the encroachment of the channel stop dopiness below the sides of the gate edge parallel to the L direction. The encroachment causes the edges of the gate to be higher doped than the center of the gate. Since the edge is higher doped, a higher gate voltage will be needed to invert the channel. Another way to think of it is that the center will have more current than the edge with a gate bias; therefore a higher gate bias will be needed to get the same effective current through the channel. The second effect is more serious than the first effect especially when higher doped channel-stop is used [10].

The last effect that will be discussed for W scaling decreases the threshold voltage. Since the threshold voltage decreases it is often referred to as the inverse narrow-width effect. The third effect occurs when the silicon is totally removed next to the gate in the L direction. The other two effects can not happen since there is no silicon to deplete. When the silicon is removed next to the gate and filled with a dielectric, it is called shallow trench isolation (STI) [10]. Please refer Figure 8 below.

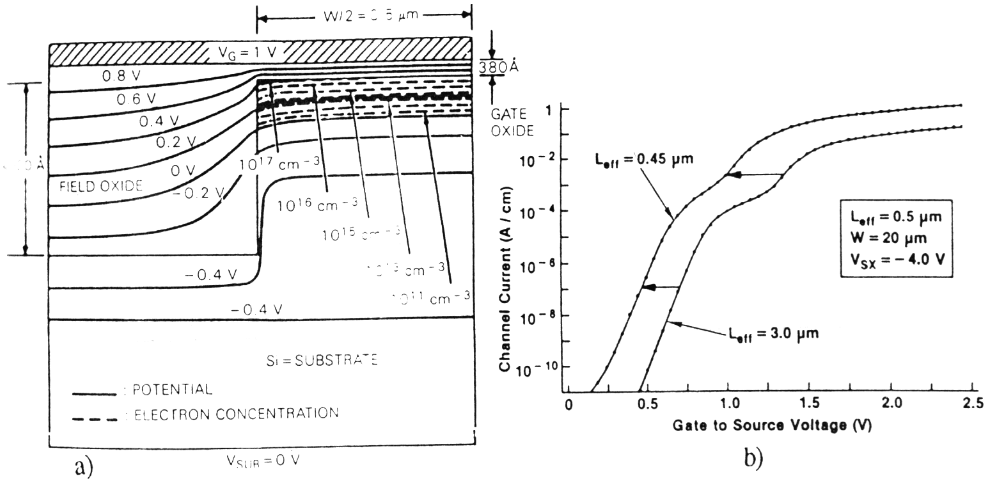


Figure 8. (a) Contours of equipotentials and electron concentrations for an STI processed MOSFET. (b) I-V plot of the inverse narrow-width effect, which illustrates the hump in the subthreshold slope [10].

Figure 8 shows the potential bending in the field oxide. Since the electric field bends at the edge of the gate this causes a concentration of more electrons to gather at the edge. The net effect causes more current on the edge than the center of the transistor. The edge region turns on sooner than the center region. The result is a lower threshold voltage. The transistor can be considered as two transistors in parallel. The parasitic (at the edge) transistor turns on before the bulk transistor. An I-V sweep shows the two transistors in which the "hump" is caused by the parasitic transistor [10].

A way of decreasing the parasitic corner Vt is by rounding the corner of the STI (see Figure 9). As the radius increases the electric field has less effect on the corner region. The effect would be similar to a birds beak [23].

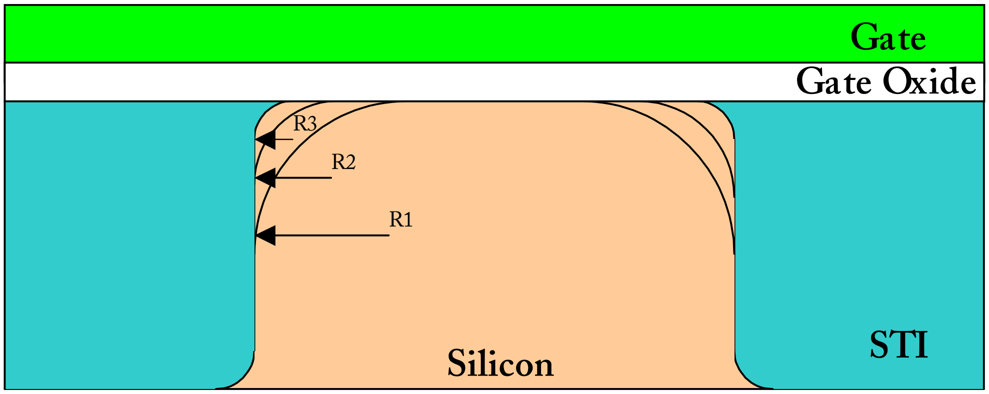


Figure 9. Corner rounding of STI to reduce the parasitic corner effect.

4. Reverse Short Channel Effects

The previous section focussed on how the threshold voltage decreased as the gate length decreased. It also discussed how width scaling could sometimes have a reverse effect on the threshold voltage. When the threshold voltage increases with scaling this is referred as Vt roll-up. The two effects, Vt roll-up and Vt roll-off, compete with one another untill Vt roll-off becomes the dominate effect as scaling increases. In Figure 10 below the two effects competing create the "hump" in the Vt vs. L_{eff} plot [10].

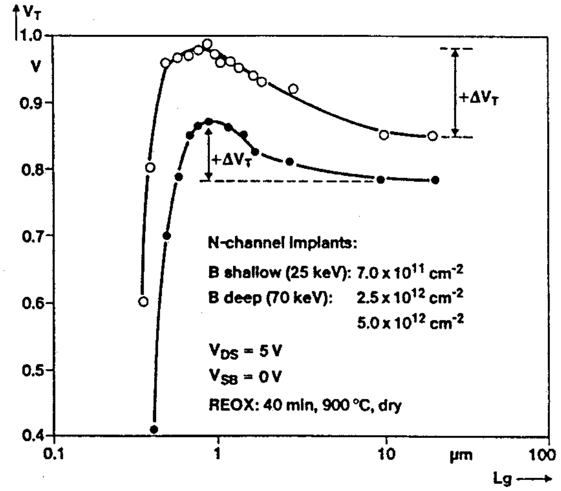


Figure 10. Reverse short channel effects (Vt roll-up) in an n-channel MOSFET [10].

5. Punch Through

Punch through occurs when the depletion regions of the source and drain meet. When the depletion regions intersect, as shown in Figure 11, the space-charge-limited current flows between the drain and source. This current cannot be controlled by the gate bias [24]. Punch through depends on the drain bias and also the substrate doping. Decreasing the drain bias will decrease the depletion region. For channel lengths below 0.1 µm, the substrate requires a doping level of 1e18 – 5e18 cm^{-3} to prevent punch through. This high doping would cause an increase in the tunneling current between the source and drain p-n junctions with the substrate [18].

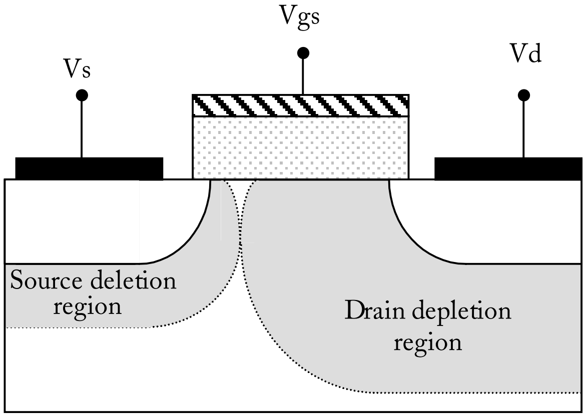


Figure 11. Schematic diagram for punch through.

6. Gate Leakage
a. Quantum Mechanical Tunneling

The leakage current when a transistor is off is very import to minimize for future transistors. The higher the leakage current the more power a chip will consume. Laptops that use batteries will lose their charge faster if the transistors have high leakage in the off state. Also DRAM retention time is dependent on I_{off}. Decreasing the gate oxide thickness proportional to the gate length helps reduce short channel effects [3]. Decreasing the gate oxide thickness helps control the electrostatic potential distribution inside the channel area [25]. As the gate length is decreased below 0.1 μm, the gate oxide thickness needs to be less than 30 Å. With ultra thin gate oxides the quantum mechanical tunneling will increase. The leakage current is exponentially related to the gate oxide thickness. Different oxide thicknesses have been grown to determine the gate oxide tunneling current. The results of the experiment suggest that tunneling current through the gate oxide will not be the limiting device leakage current for gate oxides as thin as 20-25 Å. The experiment was conducted to determine the power consumption of logic chips. Other effects will need to be determined for ultra thin gate oxides, such as reliability and device yield [18].

b. GIDL

Another type of current leakage that should be considered as the gate oxide thickness is decreased is GIDL (Gate Induced Drain Leakage). When the gate is in the off state and the drain voltage is positive for an n-channel MOSFET, the electric field from the drain to the gate can cause the overlap region to form a depletion region (see Figure 12). If the electric field is high enough, the depletion region near the surface may invert to p-type. When the minority carriers are drawn to create the inversion layer, they are swept into the p-well [10]. Electrons from the valence band can tunnel into the drain region under the overlap. The holes left in the valence band drift to the p-well. GIDL does not increase because of scaling the gate length, but does increase when the oxide thickness is reduced, since the electric field increases. One way to decrease GIDL is to decrease Vds. GIDL is independent of temperature [22]. The lack of temperature dependence is a way to detect GIDL since electrical measurements of leakage can be performed at different temperatures. If the leakage current stays relatively the same for different temperatures, it is probably caused by GIDL. Also bird's beak can reduce GIDL since the electric field at the corner of the device is reduced.

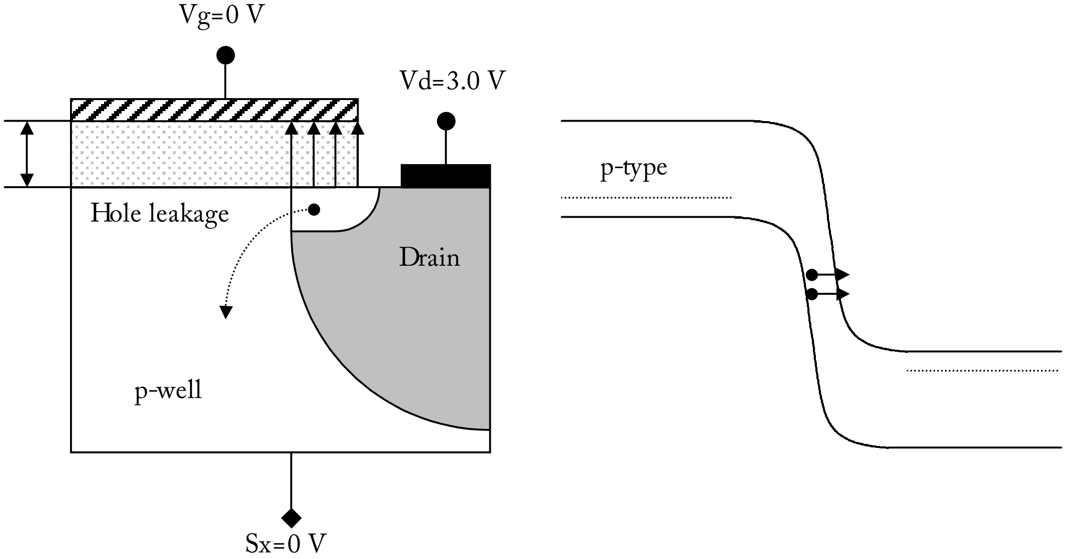


Figure 12. Schematic diagram illustrating GIDL.

7. Mobility

Mobility is a measure of the ease with which an electron or hole can move in a semiconductor. For long channel devices, mobility is determined by impurity and lattice scattering in the Si [10]. The mobility in the channel will be less for the short channel devices. One reason for the decrease in mobility is because of the effect discussed in section III.B.2., velocity saturation, which occurs as a result of the electric field perpendicular to the gate, E_y. The other electric field to consider is the one perpendicular to the channel, E_x. This electric field component causes scattering of the electrons near the Si surface. The increase in scattering slows the electrons down, thereby decreasing the mobility with respect to the bulk [15]. E_x attracts the electrons to the interface between the Si and SiO_2, and since the interface is not smooth, it will cause more electron scattering [10]. The mobility will also decrease for an increase in substrate doping.

8. Latch-Up

Latch-up in CMOS is defined when a low resistance path is created from parasitic pnp and npn bipolar transistors from V_{DD} to ground. The bipolar transistors form a Si-controlled rectifier that has positive feedback. The rectifier can form a virtual short between the power supply and ground. Figure 13 shows the npn and pnp bipolar transistors formed from a CMOS inverter cross section. The excessive current, if not stopped, can destroy the circuit or the circuit will not work properly until the circuit gains control [15]. In order to prevent latch-up, circuits must be designed so that the parasitic rectifier stays in the high impedance state [26].

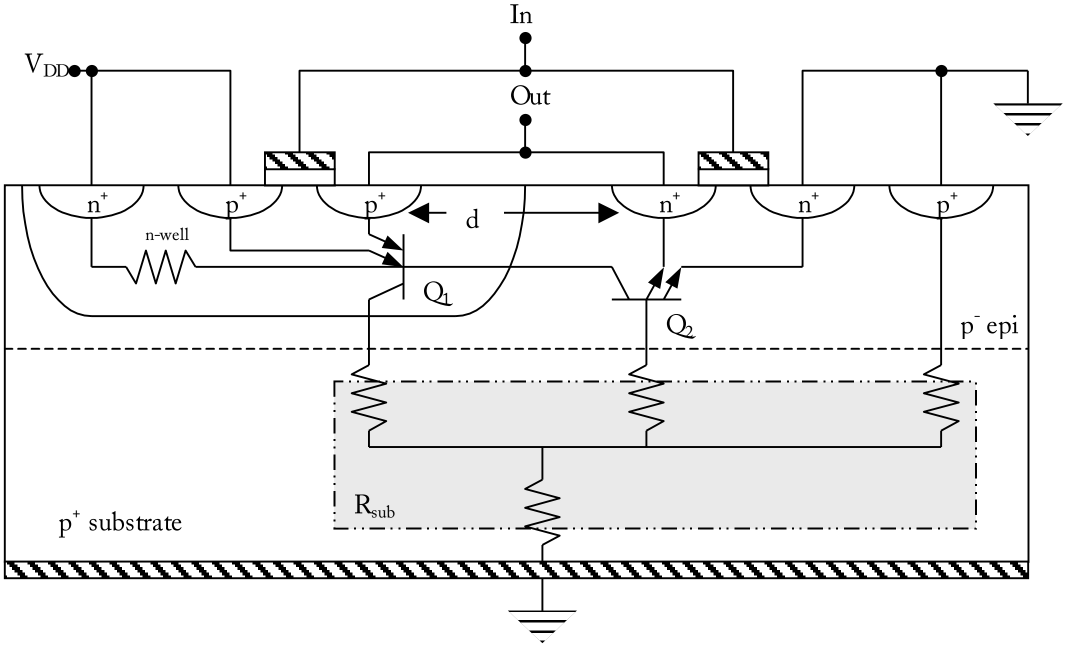


Figure 13. Schematic cross section for latch-up.

The current-voltage characteristics can be seen Figure 14 which shows the high and low impedance states. The curve has two main points (V_s, I_s) and (V_h, I_h). The current below the point (V_s, I_s) is the high impedance region and the current above is the negative differential resistance region. The negative differential resistance region continues until the second point of interest (V_h, I_h). After this point the device is in the low impedance region. V_h and I_h are referred to as the holding voltage and current, respectively, and V_s and I_s are referred to as the switching voltage and current, respectively. For most cases a good circuit for latch-up is defined when V_{DD} is less than V_s. Keeping V_{DD} below the V_s will ensure latch-up cannot continue after the transient trigger pulse becomes quiescent. If V_{DD} is greater than $V_{s,}$ the circuit could remain in latch-up after the transient trigger pulse is no longer present [26].

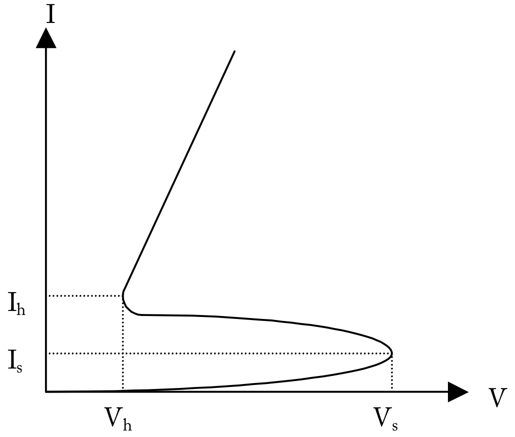


Figure 14. Latch-up I-V characteristic.

The potential for latch-up will still remain as circuit dimensions are decreased. The probability of latch-up increases as the distance d in Figure 13 decreases. IBM Microelectronics has shown that as the n^+ to p^+ distance is decreased, V_h will decrease linearly until 1 um. Below 1 μm, V_h decreases less and begins to saturate. Also V_h is dependent on trench isolation. V_h will be less than V_{DD} for most general cases that impose possible latch–up. V_{DD} should decrease as the n^+ to p^+ distance is decreased, which will help to reduce the probability of latch-up. Latch-up from the power supply will not be the main contributor from scaling effects. As devices continue to be scaled the main causes of latch-up will come from transmission line reflection from the output pads or noise coupling. Layout ground rules will force designers to consider latch-up effects in their designs in the future [26].

9. Interconnects

Interconnects can be scaled using the constant electric field scaling mentioned in section III.A. The width, length, insulator thickness, and spacing between lines can all be scaled by the constant K. The material properties are assumed to remain constant, such as the resistivity of the metal and dielectric constant for the insulator. If these assumptions are correct, the capacitance of the wire per unit length will remain the same. The wire resistance, on the other hand, does not decrease, but increases by K. The resistance per unit length (R_w) increases by K^2. Therefore the RC time constant increases by K^2. The RC time delay (τ_ω) formula is given by

$$\tau_\omega = \frac{1}{2} R_w \left(K^2\right) C_w \left(\frac{\ell}{K}\right)^2 \qquad \textbf{Equation (10)}$$

The RC delay remains constant since the K terms cancel. For aluminum (Al) this does not pose a problem since the RC delay becomes [27]

$$\tau_\omega \approx 3 * 10^{-18} \frac{\ell^2}{A} \qquad \textbf{Equation (11)}$$

which equals 1ps of delay or less. This number is much smaller than the intrinsic delay for 0.1 μm CMOS technology, which is approximately 20 ps [27,28]. Also worth mentioning is that the current density increases by K (see Table IV), forcing long term reliability issues such as electromigration to be addressed.

The above discussion is for local wires. While the RC delay time for local wires will not cause problems, the delay time for global wires will. Global wires are on the order of the chip size. They are not scaled down by K. The chip size usually does not decrease, but will more likely increase slightly as more and more transistors are added with each iteration of more powerful chips. Since the chip size is basically the same, the RC delay time will increase by K^2 for global wires. Global wires will only cause a problem if they decrease by K for each shrink. A solution for this problem is to use constant scaling for the local wires and not to scale or scale-up the global wires. Eventually the scale-up approach will have problems as it approaches the limits when the inductive effect out weighs the resistive effect. When this happens the signal rise time is shorter than the time it takes the signal to travel to the end point [27]. An alternative solution is to replace Al with a lower resistivity material, such as copper (Cu). We discuss this approach in section III.E.3.

C. Reliability Concerns in Scaled-down MOSFET Technologies

The time-zero, or initial, device characteristics are certainly of great importance and require special attention in the development of advanced MOSFET technologies. The choice of the channel length, gate oxide thickness, substrate doping and source/drain engineering determine to a great extent the device performance. When designing smaller devices, one must also consider the impact of scaled-down dimensions on the reliability of integrated circuits. Reliability engineering is concerned with how well an integrated circuit performs over time, and it is the responsibility of the reliability engineer (through modeling and accelerated testing) to ensure that the lifetime of a scaled-down device is acceptable. There are many potential failure mechanisms in modern Si MOSFET technologies, such as hot carrier degradation, gate oxide breakdown and interconnect failure due to electromigration. Generally, these mechanisms are more likely to lead to failure in scaled-down technologies because of higher electric fields and current densities. An optimum device and circuit design is one that meets both performance and reliability specifications.

1. Hot Carrier Degradation

The mechanism of hot carrier degradation in an n-channel MOSFET (Figure 15) is typically described by the "Lucky Electron" model [29]. When a device is operated in saturation mode, electrons are injected into the drain-substrate depletion region. Since the electric field is quite high in this region, some electrons acquire enough energy to cause impact ionization (i.e., electron-hole pair generation) and are referred to as hot electrons [30]. The maximum electric field is located between the pitch-off point and the drain-substrate metallurgical junction. Electrons generated in the drain-substrate depletion region may be redirected (i.e., momentum changed) toward the gate oxide. At the same time, holes generated in the drain-substrate depletion region will give rise to a substrate current, I_{sx}, as illustrated in Figure 15. (The monitoring of I_{sx} has proved invaluable as a means of modeling hot carrier effects. A higher value of I_{sx} corresponds to a higher impact ionization rate.) If hot electrons with energy > 3.2 eV overcome the potential barrier between Si and SiO_2, they may be trapped in the oxide and give rise to a gate current [30,31]. It is also possible for hot electrons with energy > 3.7 eV to generate interface traps, or surface states, at the Si-SiO_2 interface [29].

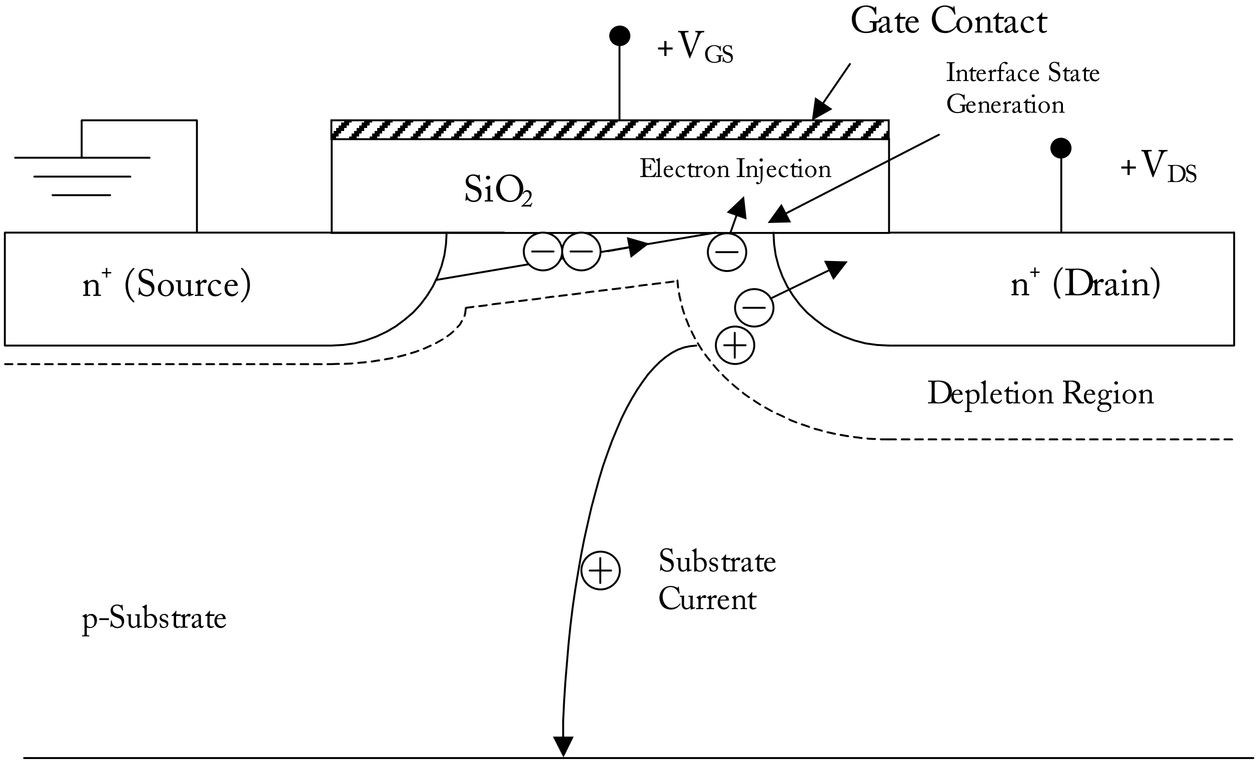


Figure 15. The mechanism of hot carrier degradation in an n-channel MOSFET.

Hot carrier degradation in n-channel MOSFETs results from interface state generation and fixed charge formation. This damage produces shifts in 1) threshold voltage, 2) mobility, 3) sub-threshold current swing, and 4) transconductance [29]. In the case of 1), the threshold voltage increases with time while the device is operated in saturation mode, which results in a lower On Current (I_{on}). In the case of 2), the mobility decreases with time, which also results in a lower I_{on}. Based on the model proposed by Hu, et al. [29], the threshold voltage increases because the generation of interface traps reduces carrier density and mobility at the drain side of the channel. The worst case device degradation is observed when I_{sx} is a maximum. Both the threshold voltage and transconductance shifts are proportional to the average trap density, which in turn is inversely proportional to L_{eff} [29,32]. Therefore, reducing the channel length will produce a lower hot carrier lifetime. (The hot carrier lifetime is defined as the time required to cause a certain threshold voltage shift or a corresponding decrease in I_{on}.) Increasing the drain-source voltage also produces a lower lifetime since the electric field in the drain-substrate depletion region is higher.

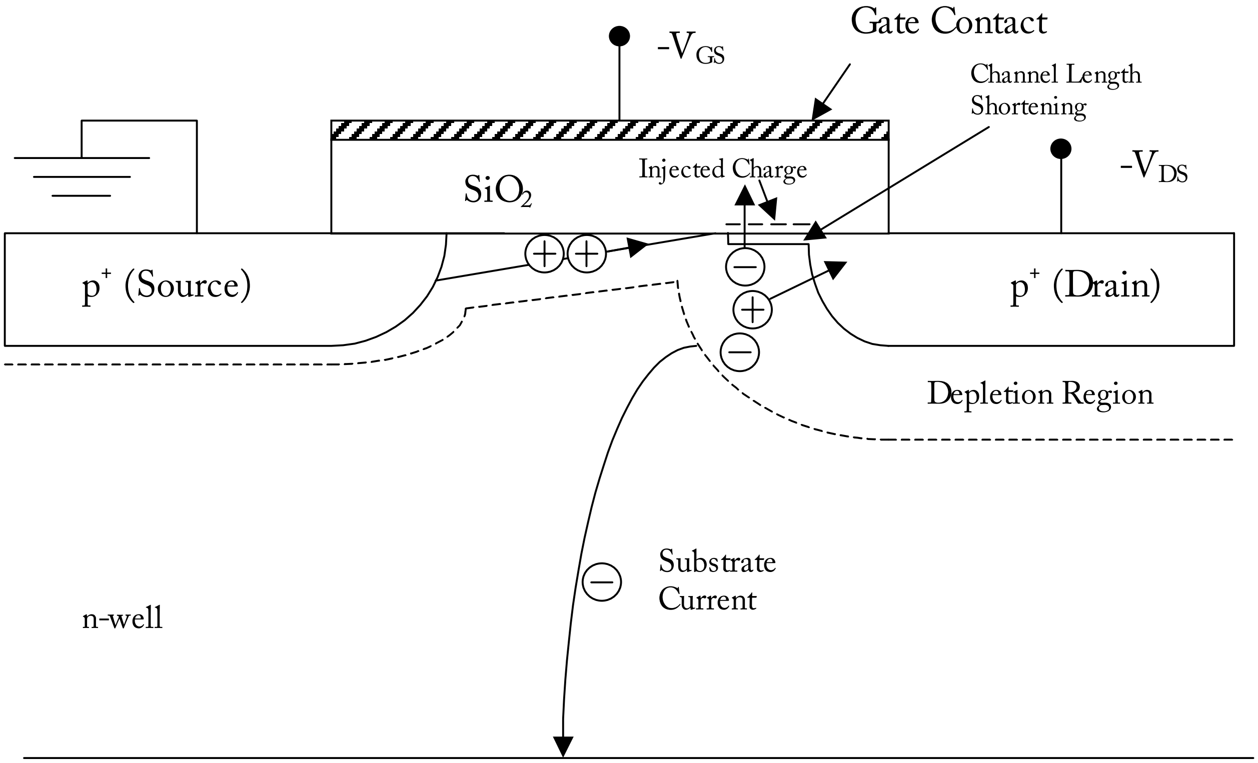


Figure 16. The mechansim of hot carrier degradation in a p-channel MOSFET.

Since the energy barrier between the Si and SiO_2 is lower for electron injection (3.2 eV) than for hole injection (4.7 eV), hot carrier degradation is more severe for n-channel MOSFETs than for p-channel MOSFETs [10,31]. Nevertheless, hot carrier degradation in submicron p-channel MOSFETs can be a serious concern [33]. When the device is operated in saturation mode, holes are injected into the drain-substrate depletion region (see Figure 16). Some holes acquire enough energy to cause impact ionization and are referred to as hot holes. Electrons generated in the drain-substrate depletion region may be redirected toward and trapped in the oxide. If the density of trapped electrons is sufficiently high, the excess negative oxide charge will attract holes to the Si-SiO_2 interface and cause an extension of the drain into the n-well region. This results in a reduction in L_{eff} and a decrease in the absolute value of the threshold voltage, |Vt|. This can be a serious problem for short channel devices, especially those that are sensitive to subtle changes in L_{eff} due to DIBL. The worst case device degradation occurs when the gate current, I_g, is a maximum. The electron trapping mechanism is dominant for |Vg|< |Vds|, while a hole injection mechanism is dominant for |Vg|> |Vds|. Hole injection has the opposite effect as electron trapping by producing an increase in |Vt|.

Hot carrier effects can be more pronounced in short channel devices because it is usually not possible to maintain the same electric field in the scaled-down device. This is certainly the situation that arises when a constant voltage scaling approach is implemented. In order to use relatively high power supply voltages and at the same time minimize hot carrier degradation, modern MOSFET technologies commonly implement a lightly doped drain (LDD) structure [30,34]. The purpose of the lightly doped region, n^-, between the drain and the channel (Figure 17) is to shift the position of the peak electric field in the depletion region toward the drain. The magnitude of the field is also reduced [30,34], where the peak electric field exhibits a minimum value as a function of the n^- dose [35]. The net effect of the LDD structure is a reduction in I_{sx}

and the impact ionization rate, which results in a lower generation of interface states and less electron injection into the oxide.

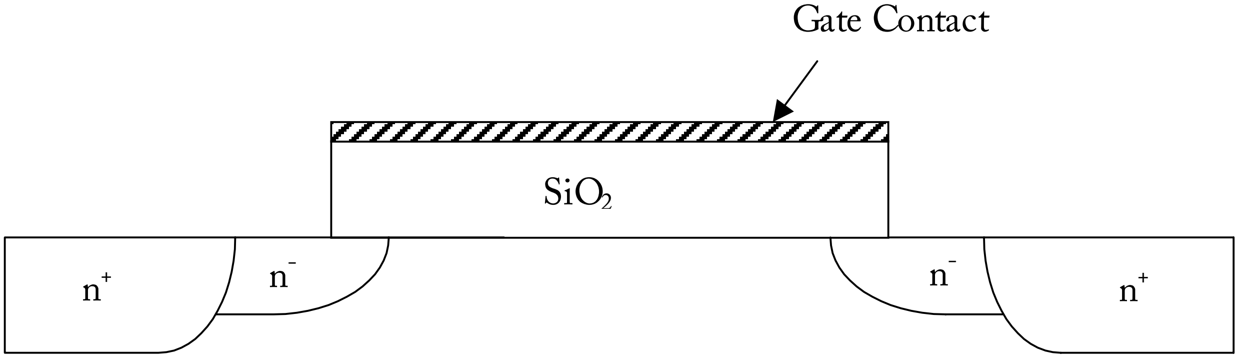


Figure 17. LDD structure for an n-channel MOSFET.

As the effective channel length approaches 0.1 μm, the gate oxide thickness reaches 3 nm, and the power supply voltage drops below 1.5 V, there is still debate as to whether or not degradation due to hot carriers will limit MOSFET scaling. For example, it has been suggested that scaling down the oxide thickness will not be limited by hot carrier degradation [22]. It was found that oxide films in the range of 5.6-15.6 nm exhibit comparable I_{on} shifts, suggesting that thinner oxide MOSFETs degrade less than thicker oxide MOSFETs when both sustain the same amount of hot carrier damage [22]. Similarly, Frey speculates that if devices become short enough, the electrons may not undergo many scattering events as they travel from the source to the drain. Therefore, the energy gained by the electrons as they arrive at the drain will be reduced as the channel length decreases, which implies that the hot carrier reliability might improve for short channel devices. [36]. Recent studies, however, have shown that hot carrier degradation is expected even at a relatively low drain-source voltage of 0.7 V [1]. It was found that the impact ionization rate is only a function of the lateral electric field, even for an effective channel length of 0.1 μm. Moreover, theoretical calculations and experimental evidence indicate that hot carrier reliability problems will persist below 0.1 μm (even as power supply voltages are reduced) due to new mechanisms such as electron-electron interactions [37,38,39,40,41] and secondary impact ionization [42,43]. These new mechanisms arise because of larger vertical fields in short channel length devices. The vertical fields are controlled by the abruptness of the drain-substrate depletion region. As L_{eff} decreases shallower junctions are required to reduce punch through effects, which results in more junction abruptness and larger vertical fields.

In the case of electron-electron interactions, it is possible for one channel electron to collide with another channel electron of the same energy. One of the electrons may lose its energy to the other electron, giving this electron two times the energy of the drain-source supply energy [37]. Simulation techniques have predicted that the high energy tail of the electron energy distribution will be dominated by electron-electron scattering for drain voltages < 3 V [41]. In one study [37], the electron energy distribution was determined by solving the one-dimensional spatially dependent Boltzmann transport equation that includes electron-electron interactions. It was found that for a long (0.25 μm) channel length device and a drain voltage of 1.5 V, the high energy tail (i.e., low probability tail) of the electron energy distribution was only slightly increased when electron-electron interactions were included. On the other hand, for a short (0.07 μm) channel length device and a drain voltage of 1.5 V, the high energy tail of the electron energy distribution was greatly increased when electron-electron interactions were considered. This implies that the hot electron population is expected to increase significantly for very short channel devices due to electron-electron interactions. Recently, Rauch, et al. [41], have shown experimentally that electron-electron scattering must be considered in order to accurately model hot carrier degradation for effective channel lengths in the range 0.07-0.10 μm.

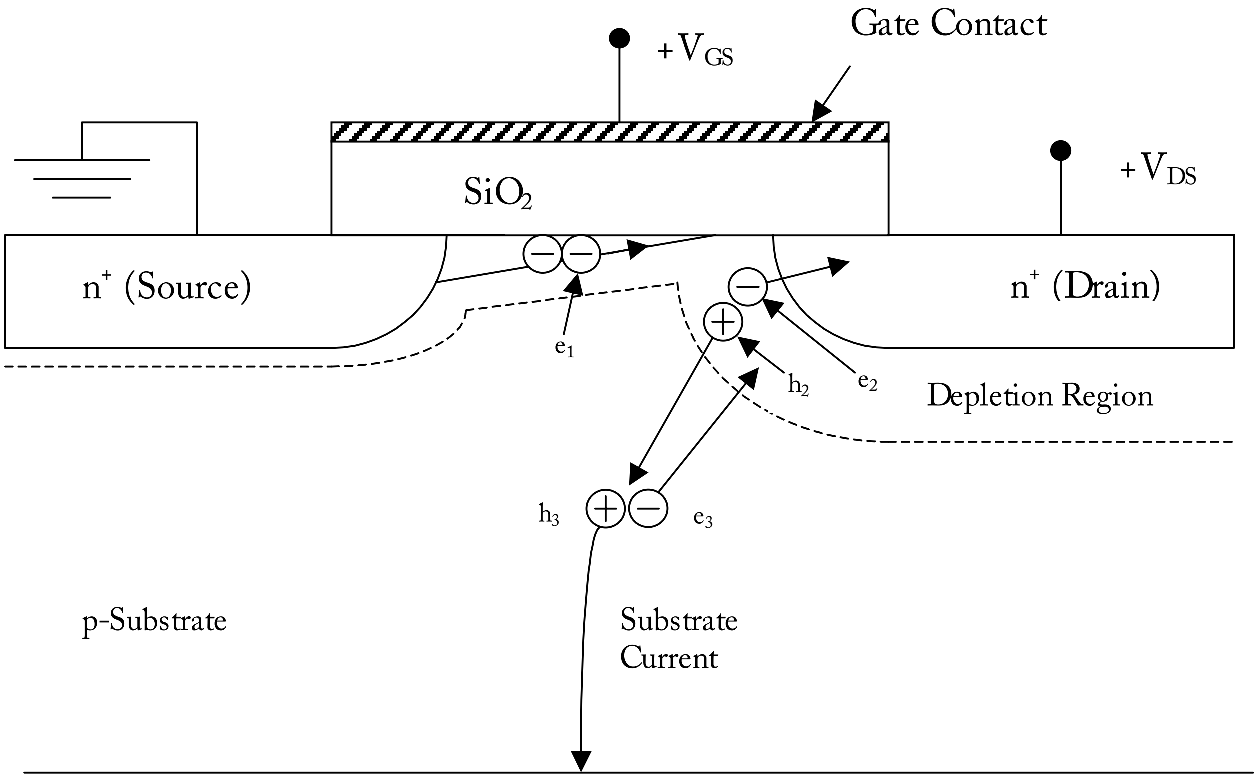


Figure 18. Impact ionization feedback process in an n-channel MOSFET.

In the case of secondary impact ionization, a feedback process occurs in which secondary electrons (holes) contribute to a gate current [43]. Figure 18 illustrates the impact ionization feedback process for an n-channel MOSFET. Channel electrons, e_1, create electron-hole pairs, e_2 and h_2, through impact ionization in the drain-substrate depletion region. The secondary electrons, e_2, are swept into drain while the secondary holes, h_2, are swept into the substrate [43]. These secondary holes create more electron-hole pairs, e_3 and h_3, through impact ionization. The h_3 holes contribute to the substrate current, while the e_3 electrons are swept back into the drain-substrate depletion region. Large vertical fields in this region can give the e_3 electrons enough energy to surmount the Si-SiO_2 barrier (3.2 eV) and thereby contribute to a gate current [43]. The e_3 electrons also lead to additional impact ionization. For relatively long channel-length devices operated at high Vds, the electron energy distribution is not affected by secondary impact ionization because of weaker vertical fields in the drain-substrate depletion region. Therefore, for long devices, the gate current will be controlled by the e_1 channel electrons. But for relatively short channel-length devices operated at low Vds, the high energy tail of the electron energy distribution is significantly affected by secondary impact ionization because of stronger vertical fields in the drain-substrate depletion region. Therefore, for short devices, the gate current is controlled by the e_3 electrons.

2. Gate Oxide Degradation and Breakdown

During device operation, the electric field across the gate oxide can be large enough for carriers to be injected into the oxide, thus causing leakage currents or catastrophic failure. In the case of leakage currents, the device may continue to operate but at a lower performance. In the case of catastrophic failure, the device ceases to operate since the oxide conducts an excessively high current and no longer has control of the charge in the channel.

In addition to carrier injection due to hot electrons, as discussed in the previous section, carriers can also be injected into the gate oxide by Fowler-Nordheim tunneling and direct tunneling [10]. (Oxide tunneling was also discussed in section III.B.6.a.) The Fowler-Nordheim situation involves electrons tunneling into the conduction band of the gate oxide (Figure 19a). This is a quantum-mechanical phenomenon in which the probability increases for larger oxide fields and thinner oxides. The gate tunneling current is exponentially dependent on the oxide field [10]. In the direct tunneling case, electrons tunnel through the energy gap of the oxide directly to the gate contact (Figure 19b). Here, there is a weaker dependence on the oxide field than for Fowler-Nordheim tunneling. Direct tunneling is dominant for oxide thicknesses < 60 nm. A lower limit on the gate oxide thickness will be reached when the tunneling current removes carriers from the channel faster than they are thermally generated [10].

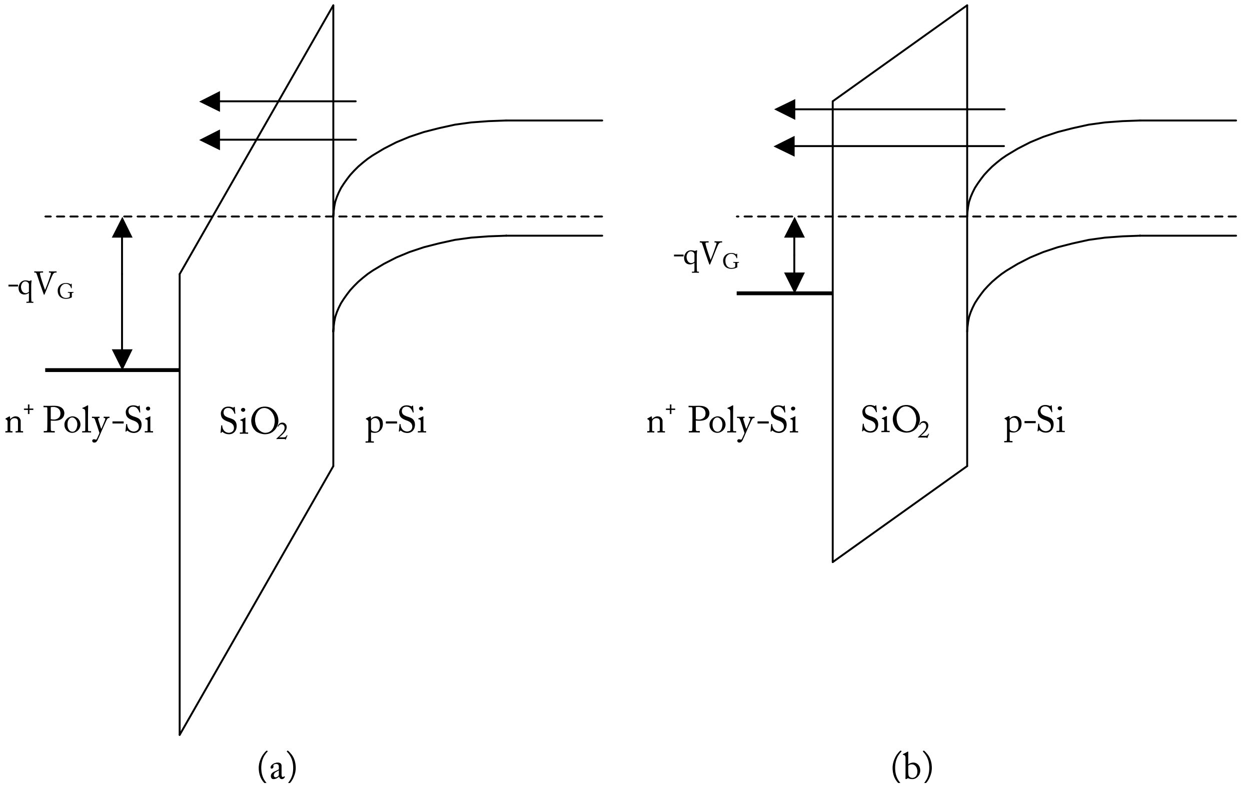


Figure 19. Schematic energy band diagrams for (a) Fowler-Nordheim tunneling and (b) direct tunneling.

Catastrophic breakdown of the gate oxide occurs at electric fields as low as 6 MV/cm and is characterized by a weakening of the oxide due to tunneling current followed by the creation of a highly conductive path between the Si substrate and the gate contact [10]. Catastrophic failures are usually placed into one of two categories; 1) extrinsic (defect-related) failures, or 2) intrinsic (wearout-related) failures. Extrinsic failures have relatively short lifetimes due to inherent weak spots or defects in the gate oxide. The defects, which include sodium contamination, metal contamination, surface roughness, and localized oxide thinning, are all related to the process of forming the gate oxide and/or to the processes that precede or follow the formation of the oxide film [10]. Intrinsic failures have relatively long lifetimes due to the absence of defects in the gate oxide. Breakdown of intrinsic oxide usually occurs at electric fields > 10 MV/cm.

Extrinsic failures ultimately represent the limiting factor for oxide reliability. The mechanism of extrinsic breakdown is not well understood, although some models have been proposed. In one model, known as the reduced barrier height model, the presence of oxide defects may reduce the barrier for Fowler-Nordheim tunneling [10,44]. In a second model, known as the oxide thinning model, defects cause oxides to act like intrinsic films with an effective oxide thickness less than the physical thickness [10]. Since extrinsic breakdown is related to defects introduced into the oxide after processing, it is important for defect levels and sizes to be properly controlled during manufacturing. A given defect level or average defect size will certainly have a greater impact on the reliability of thin oxides than on thick oxides. Defect levels can be reduced to a certain extent through adequate substrate preparation and pre-oxidation cleans [45].

Many failure mechanisms for intrinsic gate oxide breakdown have been postulated over the years [46]. In one mechanism, the interface trap density increases as the oxide degrades. This results in a local softening of the interface, which eventually leads to breakdown. In another mechanism, positive or negative charge trapped in the oxide reaches a critical value, resulting in an increase in current flow and eventual breakdown. A third mechanism is concerned with the neutral electron trap density. Here, the initial trap density is quite small but increases over time while an oxide electric field exists. Breakdown occurs when the trap density reaches a critical value and a local conductive path (through the traps) connects the gate contact with the Si substrate. Here, it is found that the critical trap density required for breakdown to occur decreases as the oxide thickness decreases. For all of these mechanisms, the probability of failure increases as the oxide electric field increases and/or as the oxide thickness decreases.

The extrinsic and intrinsic breakdown lifetimes of gate dielectrics can be improved for thin oxides (< 10 nm) by a stacked gate oxide process [45]. Here, a chemical vapor deposition (CVD) oxide film is deposited on top of the thermal oxide. The advantages of this process over a conventional thermal oxide process are: 1) lower probability that defects or weak spots in each component layer will align, 2) contains fewer defects originating from the substrate because less Si is consumed. The reliability of stacked gate oxides is significantly improved, which may allow scaling to thinner gate dielectric films than would possible for conventional thermal oxides.

For thin oxides (< 10 nm), leakage currents measured at low applied electric fields increase after application of high fields. These currents, which occur in the direct tunneling region of device operation, are commonly referred to as stress-induced leakage currents (SILCs) [46]. A trap-assisted direct tunneling mechanism has been proposed for this phenomenon, in which the generation of neutral electron traps (in the oxide) leads to an increase in current flow [47]. Very few neutral trapping sites can cause a significant increase in the direct tunneling current. Moreover, these trapping sites can be produced during oxide processing. Therefore, SILCs are expected to be of greater concern as the gate oxide thickness is scaled down. SILCs may represent a greater reliability concern than actual oxide breakdown for some MOSFET technologies, such as nonvolatile memories [46,47].

The reliability of the gate oxide may ultimately place a limit on MOSFET scaling. It has already been demonstrated that large tunneling currents (due to direct tunneling) occur in oxides thinner than 4 nm at a gate potential of only 2 V [1]. Although these large currents (10^{-3} A/cm^2 for a 3 nm oxide at 2 V) may not significantly impact the time-zero device performance, the long term effects on charge trapping and oxide breakdown may be cause for concern [1]. Hu has suggested that oxide breakdown and circuit speed will dictate the optimal choice of oxide thickness and power supply voltage for film thicknesses down to 3 nm [48]. Oxide leakage currents may limit scaling to 2.0-2.5 nm for logic applications (1 V gate voltage) and 3 nm for DRAM applications [48]. (The oxide scaling limit due to tunneling currents is a controversial subject. As mentioned in section III.B.6.a., one group speculates that tunneling current through the gate oxide will not be a limiting factor for oxides as thin as 2.0-2.5 nm [18].)

The daunting prospect of oxide reliability being the limiting factor for MOSFET scaling was emphasized at the 1998 International Electron Device Meeting. Stathis and MiMaria [49] presented data which suggest that the SiO_2 thickness cannot be reduced much below 2.6 nm (for 1 V supply voltage) due to unacceptable reliability. The slope of the breakdown distribution decreases as the gate oxide thickness is reduced, leveling off at 1 for thicknesses below 2.5 nm. This implies that failures will occur much sooner for thin oxides, thus making 10 year lifetime requirements for integrated circuits unattainable. It must be realized, however, that this study (like most oxide reliability studies) was performed on MOS capacitor structures with a critical area much less than that of actual IC chips. Gate oxide lifetime projections are based on scaling the test structure area to the actual chip area, which may not be accurate.

3. Interconnect Failure due to Electromigration

For over thirty years, ICs have relied on Al based interconnects to carry current to and from active devices. The reliability of these interconnects is generally limited by a phenomenon known as electromigration. Electromigration is the motion of atoms in a conductor due to the passage of current. It is basically a diffusion phenomenon in which momentum is transferred between the electrons and the conductor atoms [50].

Electromigration can lead to failure by one of two mechanisms. In both cases, a net amount of Al migrates in the direction of the electron flow. In the first case, a void is left behind at the negative end of the interconnect. Since the early 1980's, many ICs have utilized multilayered interconnects (i.e., Al with underlayers and/or overlayers of titanium, titanium-nitride or tungsten). Therefore, as the void continues to grow due to continued mass transport, the resistance of the interconnect increases until open circuit failure occurs [51]. In the second case, accumulation of Al occurs at the positive end of the interconnect. This accumulation causes pressure to be exerted on the surrounding insulator, and failure occurs when the extruded material reaches an adjacent interconnect.

Although electromigration performance is significantly improved by introducing small amounts of copper (Cu) into the Al matrix [52], MOSFET performance and density requirements lead to higher current densities in the metal interconnects. For a given temperature, the electromigration lifetime is proportional to j^{-n} [50], where j is the current density and n is an exponent that ranges in value from 1-2 [53]. Therefore, the interconnect lifetime is expected to decrease with each succeeding MOSFET generation (see Table IV). Cu based interconnects show a substantially longer (10-100x) electromigration lifetime as compared to Al [54,55]. Therefore, as Cu gradually replaces Al for feature sizes < 0.3 µm (primarily due to increased performance demands), failure due to electromigration is expected to be less of a concern. We discuss Cu metallization in further detail in section III.E.3.

4. Trade-off Between Performance and Reliability

The main purposes of MOSFET scaling are to 1) provide a roadmap for design and 2) require the actual design to focus on optimization of short channel effects. For digital applications, it is critical that the device threshold voltage is large enough (>0.4 V) in order to reduce I_{off} and the noise sensitivity. While scaling certainly improves performance by producing a larger I_{on} and increasing the switching speed, scaling also introduces short channel effects in submicron devices that adversely effect the device performance and long term reliability. Device designs that improve circuit performance may be detrimental to the MOSFET reliability, and vice versa. For example, reducing the depth of the source and drain regions leads to less threshold voltage reduction due to DIBL and less junction capacitance. But at the same time, shallower junctions are more abrupt and result in larger electric fields in the drain-substrate depletion region. These larger fields may cause more oxide damage than expected due to new hot carrier effects, such as secondary impact ionization [42]. Likewise, although reducing the oxide thickness improves performance since $I_{on} \propto C_{ox,}$ thinner oxides are more susceptible to catastrophic breakdown, direct tunneling currents and SILCs [45,46,47]. Conversely, while the LDD structure described above reduces hot carrier effects and allows shorter devices to be operated at higher voltages, the n^- regions increase the series resistance and thus lead to a lower I_{on} [30,34].

It is interesting to note that the introduction of Cu metallization into CMOS technologies is one area that will likely lead to increased chip performance and improved reliability [54,55]. As mentioned in section III.B.9., the RC time delay for global wires poses a problem for Al interconnects. Cu has a significantly lower resistivity than Al (1.7 µΩ-cm versus 2.7 µΩ-cm), which results in a lower interconnect RC time constant. This advantage should allow future interconnect scaling to be consistent with high performance and high density requirements. At the same time, Cu films exhibit several orders of magnitude higher electromigration lifetime compared to Al films. This should improve interconnect reliability and allow higher current densities to be used at operating conditions.

D. Techniques to Control Short Channel Effects

As Si MOSFET technologies approach the 0.1 μm regime, short channel effects such as DIBL, shifts in Vt, punch through and mobility degradation are more likely to hinder device scaling [56]. Punch through, for example, is controlled in conventional MOSFETs by using higher substrate (channel) doping as the channel length is reduced. But this approach has limitations since higher substrate (channel) doping decreases channel mobility and results in lower Vt control [56]. In order to reduce the channel length and at the same time maintain an acceptable threshold voltage, channel mobility and punch through control, state-of-the-art MOSFET technologies implement source/drain-engineered and channel-engineered devices (see Figure 20). In this section, we discuss the device characteristics and hot carrier reliability of these engineered structures.

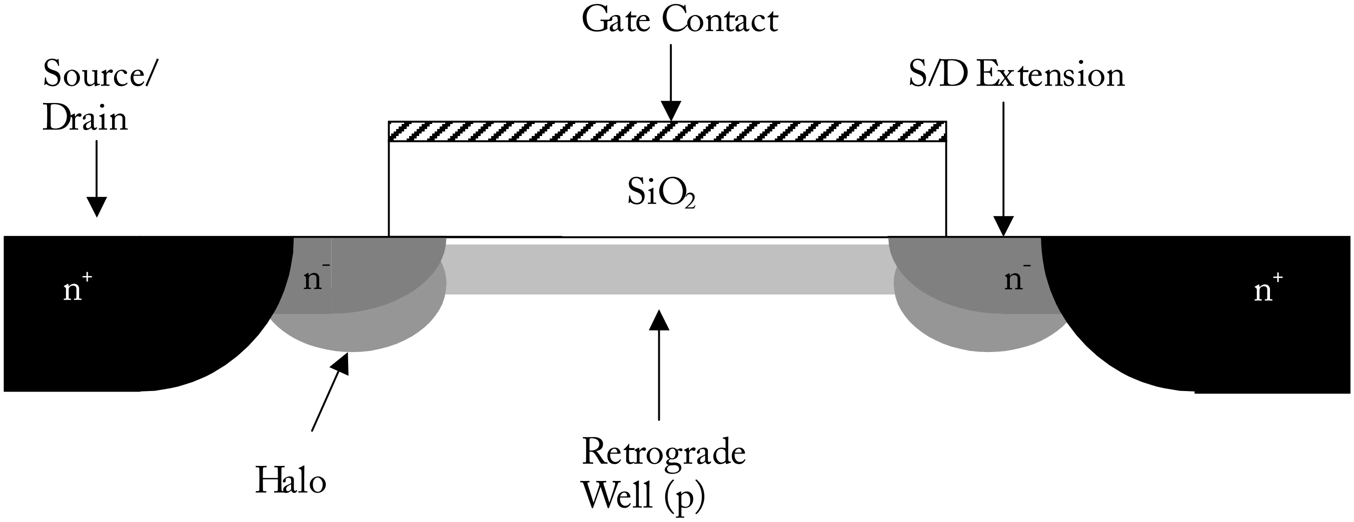


Figure 20. Schematic diagram illustrating various aspects of device engineering.

1. Source/Drain Engineering

One way to reduce short channel effects is by using source drain extensions (SDE). The LDD structure mentioned in section III.C.1. is an example of source drain extensions. SDE can be formed by first etching the gate followed by ion implantation forming the SDE. After the SDE implant a spacer will be added attached to the gate. The spacer's purpose is to block the higher dose source/drain (S/D) implants. The SDE should be relatively shallow compared to the S/D implants. The deeper the SDE the more short channel effects increase. But on the other hand, the shallower the SDE the higher the external resistance. The external resistance can be broken down to five resistors in series. The current in the channel flows first through the channel (accumulation region) next to SDE (spreading resistance) then through the deep source implant (shunt resistance) and finally through contact resistance. The main components of the external resistance are the $R_{ACCUMULATION}$ and the $R_{SPREADING}$ resistance. When transistors are scaled the channel length becomes smaller and the SDE depth becomes narrower. The channel resistance decreases but the SDE resistance increases. Scaling of the depth can not continue forever. Intel proposes that SDE depths below 30-40 nm will have little to no benefit for devices with gate lengths less than 0.1 um. The reason for this is that any gain in short channel effect because of reduced charge sharing will be balanced out because of the increase in external resistance. Also if the SDE depth is very narrow it will not extend far enough under the gate. The SDE must extend under the gate to increase drive current. If the SDE does not extend enough under the gate the current will spread out more in the lower doped part of the SDE. This will cause an increase in the $R_{ACCUMLATION}$ and $R_{SPREADING}$ resistance. The increase in the overall external resistance will decrease the maximum drive current [4].

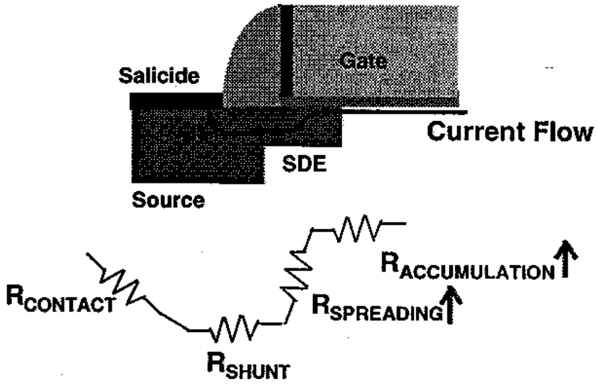


Figure 21. Components of external resistance [4].

As was presented above, the SDE is created using an ion implantation. Ion implantation can cause implant channeling and also cause transient enhanced diffusion. Both of these effects can cause the SDE to be deeper than intended, which will increase short channel effects. Decreasing the implant energy can control the effect, but this will not be able to continue for smaller SDE for p-channel MOSFETs using Boron as the implant. An alternate method for doping the channel is to use BSG (borosilicate glass) indiffusion. The BSG is place directly on the silicon, as illustrated in Figure 22. This process has been shown not to suffer from the same effects [57].

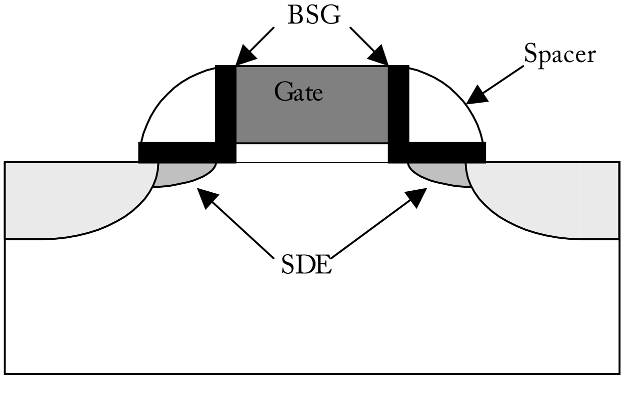


Figure 22. BSG (borosilicate glass) indiffusion

2. Channel Engineering

a. Lateral Channel Engineering - Halo Implants

One way of reducing short channel effects by extending Vt roll-off is through halo implants. Halo was originally observed in oxidation-enhanced-diffusion that is challenging to control [16]. Halo implants increase the doping near the source and drain implant (see Figure 23).

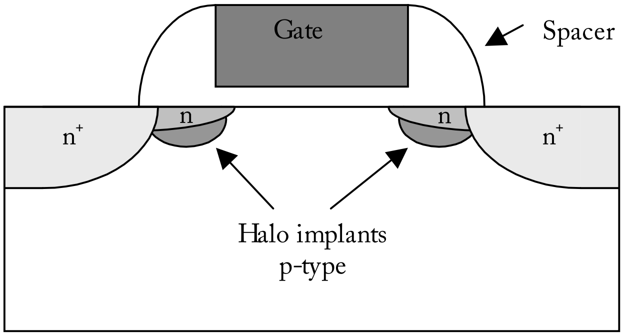


Figure 23. Halo implant for an n-channel MOSFET.

The implants near the source or drain can be symmetrical or asymmetrical. Halo implants do help reduce short channel effects such as DIBL, punch through, and Vt roll-off. But also halo reduces drive current in the transistor. The trade off between drive current and the reduction of short channel effects must be considered to maximize performance of the transistor [58]. The halo implants can be vertical or can have a tilt. They are usually added after the gate pattern is finished. The implants add more of a barrier between the source drain junction with the channel.

Halo implants only add small increases of Vt for long devices, but much larger increases can be seen with short channels. Figure 24 below shows the doping concentration for a halo and non-halo device.

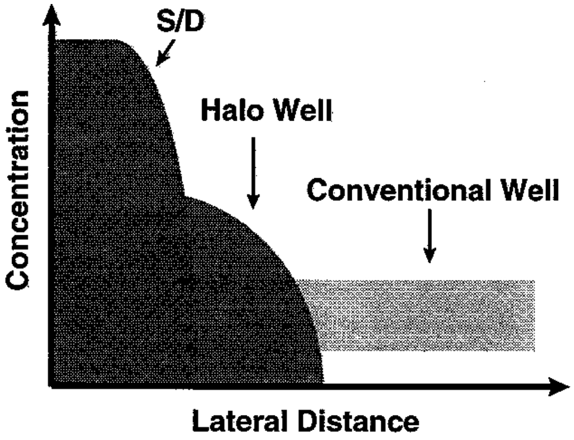


Figure 24. Schematic showing a lateral surface cut of the well doping near the SiO_2 interface [4].

For conventional MOSFETs the threshold voltage most not roll off more than 100 mV. According to this requirement, transistors could not be fabricated below 0.3 um without the use of halo implants. With halo implants, transistors can be scaled well below 0.2 um [59]. Figure 25(a) below shows the increase in the threshold voltage as the halo implant is increased. Figure 25(b) after that shows the decrease in Ioff for devices with halo implants. Also, Figure 25(a) shows Vt roll-up which is more dominant for the strong halo.

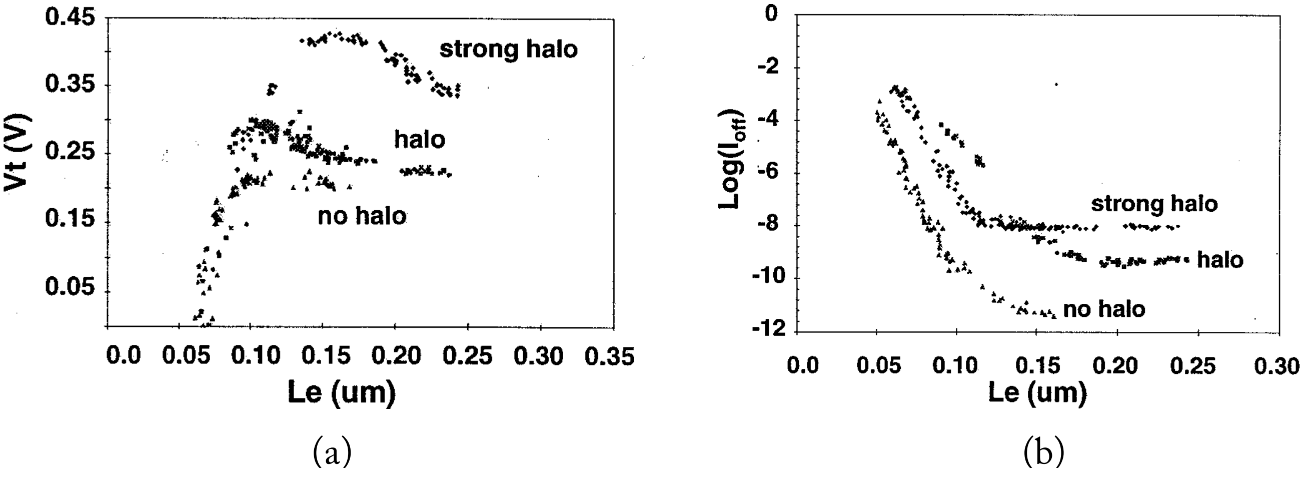


Figure 25. n-channel MOSFET characteristics for no halo, halo and strong halo. (a) Vt versus L_{eff} and (b) Log(I_{off}) versus Leff [4].

IBM reported on 0.08 um MOSFETs with strong halo to control short channel effects. They found the halo implant strongly effects the I-V curves for an intrinsic n-channel MOSFET. Intrinsic MOSFETs are used in analog PPL designs and also used in I/O voltage translation circuits. The Vt is increased but as the gate voltage increases the I-V curve changes slope as if it was a long channel device with a very low Vt. The MOSFET can be interpreted as a short channel MOSFET with a high Vt in series with a longer channel with a low Vt. Figure 26(a) below is a simulation done to compare MOSFETs with and without halo implant. As can be seen in the figure there are two linear regions for the transistor with a halo implant. The first linear region is the halo region and the second linear region to the right is the low Vt transistor. Figure 26(b) plots actual measured data from an intrinsic transistor [60]. The figure shows that Vt increases as the substrate is biased more negative.

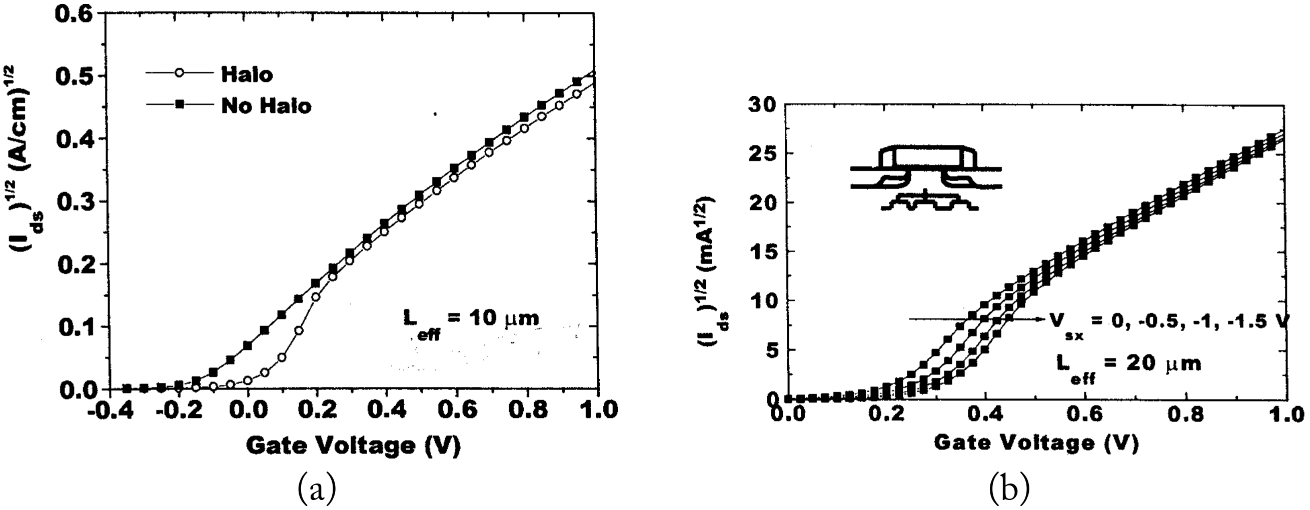


Figure 26. (a) Simulated I-V with and without halo and (b) Actual I-V with halo and substrate bias [60].

The Amercan Institute of Physics reported how halo implants shift the Vt of long devices (20 μm width, 20 μm length). The halo implant should have little effect of longer devices. The report found the reason the longer devices Vt shift is because the halo implant channels through the polycrystalline silicon gate. The conclusion of the paper was to either increase cap oxide over the gate to reduce the channeling or reduce the implant dose [59].

b. Vertical Substrate Engineering - Retrograde Channel Profiles

Conventional channels, formed by implanting dopants into the substrate and diffusing them (at high temperature) to a certain depth, can be quite susceptible to short channel effects such as punch through, mobility degradation and latchup. One design that minimizes these undesirable effects is known as a retrograde profile. Such a design is achieved by using high energy ion implantation to place dopants at a desired substrate depth and then annealing at a low temperature to activate the implants [10]. A key feature of retrograde structures is the use of slow diffusing dopants such as arsenic or antimony for p-channel devices and indium for n-channel devices [4]. Figure 27 illustrates the doping profiles (simulated) in conventional and retrograde implanted p-well structures. As can be seen from the figure, the doping concentration for the conventional well is highest at the Si surface and decreases as one moves further into the p-well. The peak of the retrograde doping profile, on the other hand, is highest at a certain depth within the Si substrate and decreases as one approaches the Si surface. The slope of the doping profile between the location of peak concentration and the Si surface can be quite high, as is the case for super steep retrograde (SSR) channel profiles. Some of the advantages of retrograde over conventional channel engineering are summarized as follows:

1) Increased packing density since high energy ion implantation results in less lateral diffusion of the implanted dopants [10].

2) Higher surface mobility (i.e., less impurity scattering in the channel region near the Si/SiO_2 interface) due to a lower surface concentration of dopants [56,61].

3) Better control of Vt due to lower surface doping [56].

4) Reduction in DIBL because the drain depletion width extends less into the retrograde well, resulting in shorter minimum channel lengths for the same I_{off} leakage current (see Figure 28) [4].

5) Better control of punch through since the doping concentration at the bottom of the well is higher [10].

6) Increased protection against latchup because the conductivity in the bottom of the well is increased [10].

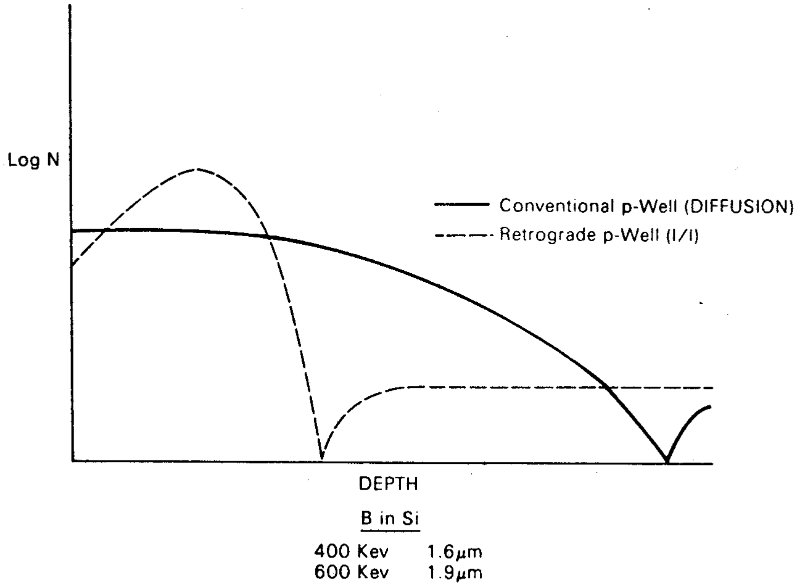


Figure 27. Comparison of conventional and retrograde p-well profiles [10].

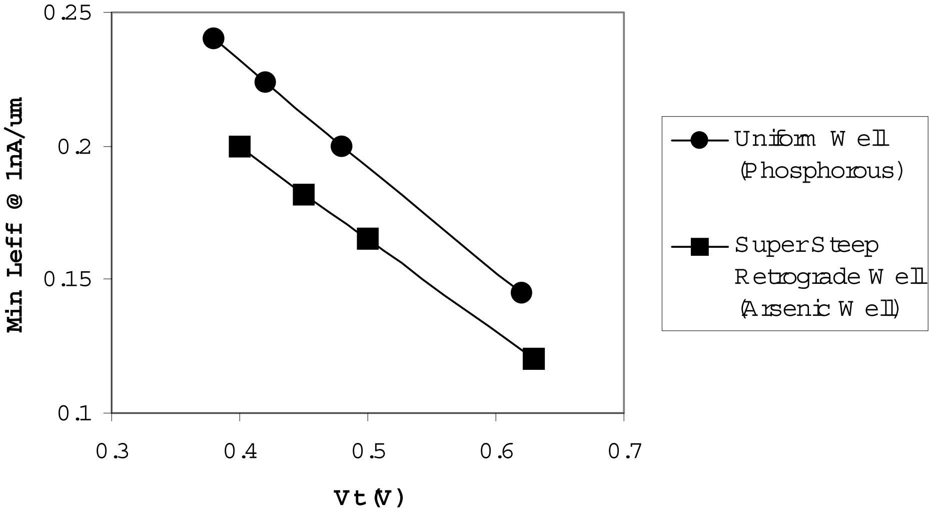


Figure 28. Minimum L_{eff} versus threshold voltage for SSR and conventional well structures [4].

While SSR channel profiles are known to improve short channel effects in sub-0.35 μm CMOS technologies [62], devices with retrograde profiles have lower saturation drive current than that of conventional profiles [62,63]. (It should be noted that the linear drive current is significantly higher for retrograde profiles compared to uniform doping profiles [63]. This result is not surprising given that the surface mobility is higher for retrograde devices.) The drive current, I_d, in saturation is given by [62,64*]*

$$I_d = \frac{W}{L}\mu C_{ox}\frac{(V_g - V_t)^2}{2(1+\Delta)}; \quad \Delta \approx \frac{\gamma}{4\sqrt{\phi_B}}$$

Equation (12)

where μ is the mobility, ϕ_B is the surface potential under strong inversion and γ is the body factor. Since γ is proportional to $(N_{SUB})^{1/2}$, where N_{SUB} is the peak doping concentration in the substrate, I_d is expected to be lower for retrograde profiles. For a 0.35 μm technology, a 10% reduction in drive current was observed for p-channel devices, while a 5% decrease was observed for n-channel devices [62]. At the same time, significant improvements in short channel effects, such as Vt roll-off and DIBL, make it possible to offset the degradation in drive current by choosing a smaller L_{eff} [62].

Although retrograde profiles help control short channel effects, such as DIBL and punch through, in modern CMOS technologies, it is not clear whether such channel-engineered devices will provide the same benefits in the deep sub-micron regime. In one study [65], simulations on n-channel devices with L_{eff} =0.14 μm indicate that the threshold voltage change due to DIBL, δVt(DIBL), is lower for retrograde profiles than for uniformly doped devices at the same threshold voltage. In addition, the effective mobility was calculated to be higher for retrograde profiles than for uniformly doped devices for a given Vt and δVt(DIBL). In a more recent study [56], however, simulations on n-channel devices with L_{eff} in the range 0.1-0.8 μm indicate that SSR designs provide less current drive and lower carrier mobility than conventional designs near the 0.1 μm regime. For longer channel lengths (0.8 μm), the SSR device shows a higher peak electron velocity than a conventional device. But for $L_{eff} \leq 0.2$ μm, SSR and conventional designs exhibit the same peak electron velocity and nearly the same electron velocity profile. This implies that the observed higher mobility in retrograde profiles will disappear as channel lengths approach 0.1 μm.

3. Hot Carrier Effects in Engineered MOSFETs

As explained above, sub-micron CMOS technologies commonly implement channel and source/drain engineered MOSFETs in order to control short channel effects. While these structures are expected to reduce the change in Vt due to DIBL and improve punch through control, the impact of device engineering on long term reliability must also be considered. Specifically, the effect of device engineering on hot carrier degradation needs to be factored into the equation for determining the optimum device design.

Since halo implants increase the doping concentration near the drain-substrate junction (or drain extension-substrate junction in the case of LDD structures), the junction becomes more abrupt and the electric field in this region increases [66]. In addition, the location of the peak electric field moves toward the Si/SiO_2 interface when halo implants are used. Therefore, it is widely believed that devices with halo implants are more susceptible to hot carrier degradation. At a drain-source bias of 5.0 V, n-channel MOSFETs (L_{eff}=0.5 µm) with LDD implants showed a higher ratio of substrate current, I_{sx}, to drain current, I_d, for higher halo doses [67]. Not surprisingly, devices with higher halo doses showed more degradation in the saturated drain current, I_{dsat}, and Vt. (For hot carrier degradation due to interface state generation, the shift in I_{dsat} is proportional to $(I_{sx}/I_d)^m$, where m>1 [29].) Also, structures with the halo implant performed after the sidewall spacer process and source/drain implantation exhibited increased hot carrier degradation compared to structures with the halo implant preceding the sidewall spacer and source/drain implantation [67]. In a recent study, the tilt angle (see Figure 29) as well as the energy of the halo implant was found to play a significant role in hot carrier behavior [66]. It was found that for n-channel MOSFETs (Leff=0.20 µm) with similar device characteristics (i.e., Vt roll-off), larger tilt angle (45°)/lower energy halo implant devices showed less hot carrier degradation than lower tilt angle (25°)/higher energy halo implant devices.

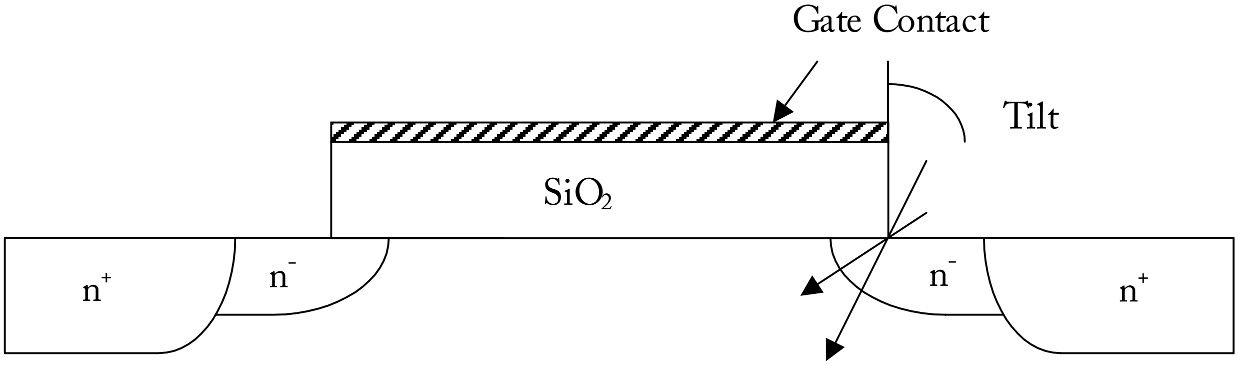


Figure 29. Halo implants with different tilt angles for an LDD n-channel MOSFET.

The effect of retrograde profiles on hot carrier induced damage has not been widely studied or modeled. Nevertheless, at least one group of investigators suggests that SSR designs with $L_{eff} \leq 0.2$ µm will exhibit greater sensitivity to interface state generation compared to conventional designs [56]. Simulations on n-channel MOSFETs indicate that the peak channel electric field is always higher (≈ 10% higher) for SSR designs than for conventional designs in the range 0.1-0.8 µm. Even though conventional designs are expected to experience more hot carrier injection into the gate oxide for Leff ≤ 0.2 µm, SSR designs should exhibit more hot carrier induced degradation due to a hotter carrier distribution [56].

As mentioned in section III.C.1., modern MOSFET technologies commonly implement LDD structures in order to shift the position of the peak electric field toward the drain and reduce the magnitude of the peak field [30,34]. Source/Drain engineered structures such as these are much less susceptible to hot carrier induced degradation. The peak electric field along the channel exhibits a minimum value as a function of the n^- dose [35,67]. Increasing the n^- dose above this point will cause the peak field to increase because L_{eff} is reduced [67]. Therefore, hot carrier degradation in LDD devices can be optimized by controlling the n^- dose.

E. Unconventional Approaches to MOSFET Scaling

Conventional approaches to MOSFET scaling continue to be the mainstream strategy for producing high performance IC chips. Reducing the effective channel length and gate oxide thickness result in faster devices, while implementation of halo implants and retrograde channels minimize short channel effects such as DIBL and mobility degradation. As CMOS technologies approach the 0.1 µm regime, however, conventional scaling techniques are expected to encounter many obstacles. For example, it is a tremendous lithography challenge to reproduce feature sizes less than or equal to 0.1 µm. Also, direct tunneling currents and gate oxide reliability concerns may prevent oxide thicknesses from being scaled below 1 nm. Finally, for features sizes < 0.25 µm, chip performance will be limited by the RC interconnect delay rather than the intrinsic gate delay. As a consequence, many companies in the semiconductor industry are developing unconventional technologies that are expected to become commonplace in sub-0.10 µm CMOS circuits. In this section, we discuss three of these emerging technologies: SOI, vertical transistor and Cu metallization/low K dielectric constant insulators.

1. SOI as an Increased Device Performance Alternative

Silicon on insulator (SOI) technology comes in response to the desire for greater circuit performance and lower operating voltages as conventional Si CMOS technology approaches its fundamental constraints. It has been discussed earlier the short channel effects (SCE) suffered by scaled MOSFET devices such as Vt roll-off and the effect on I_{off}, drain induced barrier lowering (DIBL), punch through, and hot carrier effects due to high electric fields of the drain/source p-n junctions with the substrate. Various methods of channel engineering have been applied to suppress these effects. Source drain extensions such as lightly doped drain (LDD) to reduce electric field of the source/drain p-n junction, halo implants to reduce Vt roll-off and DIBL, and retrograde channel doping to control punch-through and electric fields have been utilized with great success. Although the "fundamental" channel length limit set at various times has always been surpassed, many still concede that the limit for channel length of bulk CMOS is on the horizon and a few have posed the value at 50nm [68,69].

SOI provides excellent performance at very short channel lengths. The literature indicates devices as low as 10nm have been measured [70]. When considering the advancement of performance the Vt/Vdd ratio is often used. Ratios as low as 0.2-0.3 are required for speed, while anything lower results in greater circuit dependence on Vt variations due to processing. Taking into account the operating frequency of wireless applications of 1-2GHz, the supply voltages desired are in the range of 1-1.5V. This means that current CMOS technology targeted for use in mobile products will require threshold voltages of 200-300 mV and result in large subthreshold currents (I_{off}) [69].

a. SOI Processing

Silicon-On-Insulator is typically processed by one of two methods. The first, called SIMOX (Separation by IMplanted OXygen), involves implanting oxygen of a certain density within the silicon wafer. When heated, the oxygen activates and forms a channel of SiO_2 below the surface of the silicon. The top surface of Si is then polished to obtain the appropriate SOI thickness.

Finally, the gate oxide is processed on top to create a layer of silicon sandwiched between two layers of oxide. The bottom oxide (BOX) is typically a few hundred nm thick but this can be adjusted for concepts such as the double gate (DGFET) to be discussed later. This is currently the leading commercial production method.

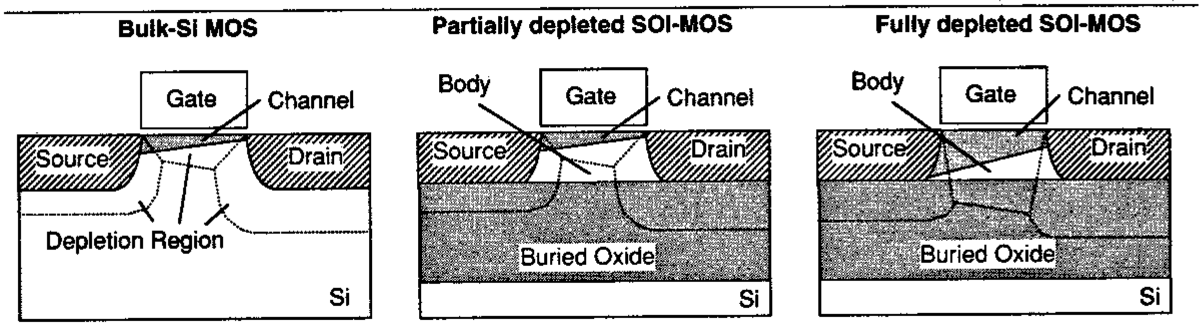
The second method of processing is called BESOI (Bonded and Etchedback SOI) and is comprised of joining an already developed oxide layer and Si layer to form a permanent bond. There are two techniques to perform this action. In the first, an epitaxial silicon layer with a buried etch stop is placed in contact with a substrate or "handle" wafer. The handle wafer already has an oxide on the surface which forms the BOX. When heated, the epitaxial layer is bonded to the handle wafer. Next, the new top Si surface is etched to the stop to form the SOI layer, and gate oxide can be placed on top. The second technique is similar to the first except that instead of using an etch stop, a hydrogen layer is implanted into the epitaxial Si layer. After joining, the compound is heated to activate the hydrogen, which forms a pocket of gas and breaks the thin layer of epitaxial silicon off from the bulk, but it remains attached to the handle wafer to form the SOI layer. The BESOI method can be used to form more complex layer geometries for special devices. In addition, this second technique can be ramped up quickly for mass production.

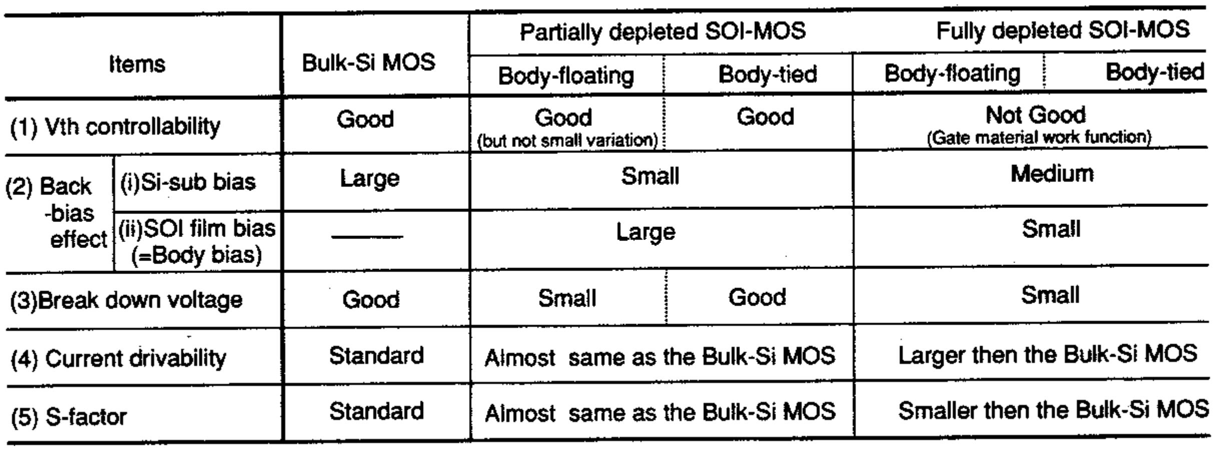
After the thin oxide and gate are placed on the SOI layer, the source/drain diffusions can be placed by self-alignment. The junction depths of the source and drain are the thickness of the SOI layer (i.e., the diffusions rest on the bottom oxide layer).

The SOI layer thickness and doping can be altered to produce two types of silicon-on-insulator called partially depleted (PD), and fully depleted (FD), SOI. The difference between the two structures of SOI devices lies in the fact that for FD devices, the doping is uniform and low enough and the SOI is so thin (less that the depletion region formed by source/drain diffusions to body), that the entire region between source and drain is depleted of majority carriers. For PD devices, there exists a region of silicon between the source and drain which is not depleted. This body can be contacted (with area penalty) or left floating depending on the desired device characteristics. The later choice can result in floating body effects (FBE) which reduce device performance. In order to reduce FBE, the channel can be adjusted in many of the same ways as bulk Si, for example, by retrograde doping and halo implants. PD vs. FD device performance will be discussed later. Table VI illustrates the difference between bulk CMOS, PD, and FD device cross-sections and their general operation characteristics [71].

Table VI. Comparison between bulk CMOS, PD SOI and FD SOI devices [71].

FEATURES OF SOI MOSFET's





Items		Bulk-Si MOS	Partially depleted SOI-MOS		Fully depleted SOI-MOS	
			Body-floating	Body-tied	Body-floating	Body-tied
(1) Vth controllability		Good	Good (but not small variation)	Good	Not Good (Gate material work function)	
(2) Back -bias effect	(i)Si-sub bias	Large	Small		Medium	
	(ii)SOI film bias (=Body bias)	———	Large		Small	
(3)Break down voltage		Good	Small	Good	Small	
(4) Current drivability		Standard	Almost same as the Bulk-Si MOS		Larger then the Bulk-Si MOS	
(5) S-factor		Standard	Almost same as the Bulk-Si MOS		Smaller then the Bulk-Si MOS	

b. SOI Device Operation

The benefits of SOI over bulk CMOS with regards to performance and power with greatly reduced SCE have already been mentioned. In particular, these advantages over bulk CMOS exist for several reasons: a decrease in the parasitic diffusion capacitances, reduced (or negligible for FD) substrate bias effect, reduced vertical gate electric field, and reduced or negligible drain/substrate electric field.

SOI speed improvements of 15-50% over traditional CMOS are commonly reported in various circuit implementations. An SOI DRAM group reports a 27% improvement in tRAC (access time) [72]. Another group reports ~30% improvement in NAND gate propagation delay depending on supply voltage [69], and others report similar improvements for such circuits as ring-oscillators and critical paths through microprocessors. These improvements are mostly due to the decrease in parasitic source and drain junction capacitance for SOI devices. Since the drain and source regions extend to the bottom of the SOI layer, no p-n junction capacitance is realized along the bottom of the diffusion. Since the bottom oxide can be made relatively thick (~1um), the capacitance with respect to the silicon substrate is very low. Comparing the diffusion capacitance for a bulk CMOS, C', vs. the capacitance seen between drain and substrate (across the BOX) for an SOI device, C_{ox}', we find:

$$C\,'=\left(\frac{qK_s\varepsilon_0 N_d}{2(V_R+\Psi_0)}\right)^{1/2}=5.78x10^3\,pF/cm^2$$ Equation (13)

$$C_{ox}\,'=\frac{K_{ox}\varepsilon_0}{t_{ox}}=3.45x10^3\,pF/cm^2$$ Equation (14)

using a body doping of 10^{15} cm^{-3}, and a BOX thickness of 1um. Here we already see a 39% improvement for the SOI just due to a reduction in parasitic capacitance.

Problems arise with the required channel engineering for reduction of SCE, particularly hot carrier effects and punch-through, in extremely short channel bulk CMOS. Retrograde channel doping results in a large ΔV_t due to processing variations of the doping concentration and channel length. The equation below relates this shift in the threshold to channel doping and geometry where L_g is gate length, w_D is depletion depth and A is a constant:

$$\Delta V_t=\frac{3A}{4}t_{ox}N_{channel}^{1/6}(L_g w_D)^{-1/2}$$ Equation (15)

Using Monte Carlo analysis for processing variations, the standard deviation of ΔV_t is 5mV for devices of 1um in length and 18mV for devices of 0.1um in length [70]. This type of variation is much too large for some critical circuitry such as differential sense amplifiers where the signal strength is on the same order. The use of FD SOI devices with uniform low channel doping eliminates this type of threshold voltage sensitivity.

Other advantages of SOI can be observed. To reduce punch-through, high doping is required for short channel bulk CMOS devices. For devices on the order L_{eff}=50nm, local substrate doping of $5x10^{18}$ cm^{-3} is required. This high doping can result in leakage current from drain to substrate (i.e., tunneling current). The low doping concentrations in the channel for SOI devices (FD in particular) eliminate this tunneling current. The existence of the BOX reduces the vertical electric field from the gate into the body of SOI structures and thus results in less mobility reduction in the channel. Finally, the existence of a conducting channel at the bottom silicon and oxide interface, as in the case of the DGFET shields parasitic field lines from drain to source which would otherwise result in hot carrier effects.

Comparing PD and FD SOI devices from Table VI, partially depleted devices have current driveabilities much closer to that of bulk CMOS. The process window is much larger than for fully depleted devices for better manufacturing. Floating body effects exist for PD devices. Fully depleted devices show a much greater current driveability and show negligible FBE but Vt controllability for such thin SOI becomes a daunting task.

Two main challenges that SOI technology faces are a short channel effect called the kink effect and FBE. The kink effect can be observed for PD SOI devices as a change in the output conductance as seen in the Id vs. Vds curves of Figure 30(a).

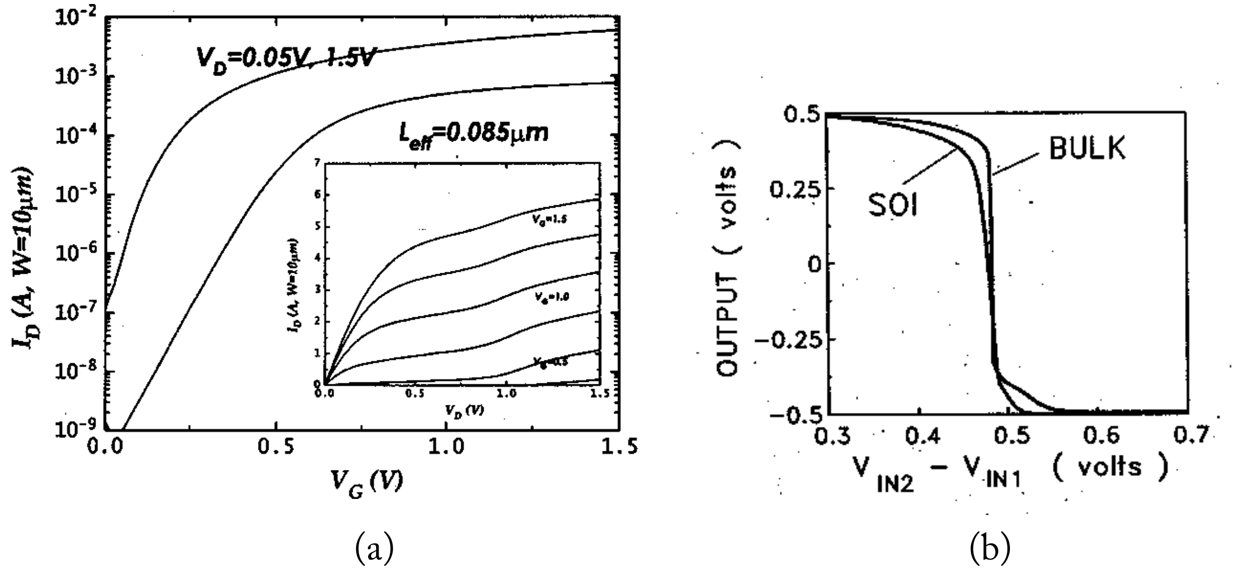


Figure 30. (a) Id v. Vds for PD SOI device [68] and (b) comparison of bulk to SOI CMOS differential amplifier transfer curve [73].

This change in output conductance is detrimental to the operation of analog differential amplifiers as it significantly reduces gain. This can be seen as a reduction in the slope of the transfer curve of the SOI circuit compared to a bulk CMOS circuit shown in Figure 30(b).

Floating body effects are manifested as a decrease in the sub-threshold Id-Vgs slope and a high I_{off}. The main strategy to eliminate FBE is to make the SOI ultra-thin in order to fully deplete the substrate (go from PD to FD device). Another alternative which is growing in popularity is to use PD devices and tie the body to a controlled voltage. This does result in an additional area required for this contact, but the control of the body in SOI structures can be used to further increase device performance. A cross section of such a device used in a 16Mb SOI DRAM is shown in Figure 31 [71].

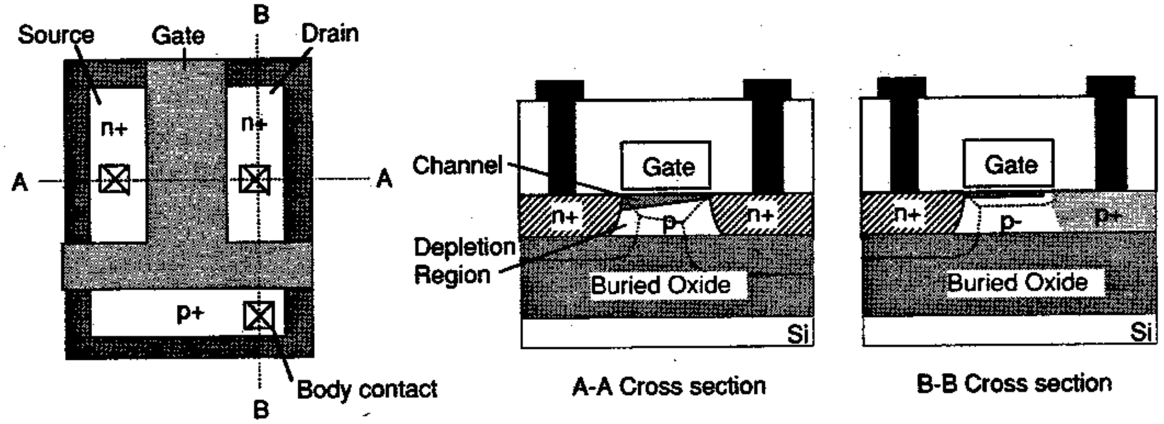
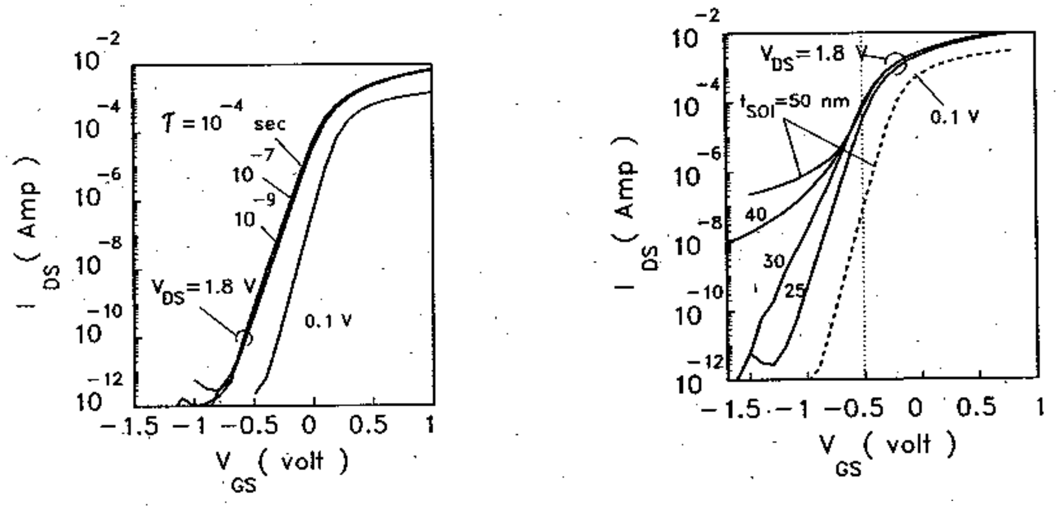


Figure 31. Cross section of a PD SOI device with body contact [71].

During periods when the device is not switching or in standby states, the body voltage can be brought low to increase threshold voltage and reduce I_{off}. When the device goes into operation, the body voltage can be increased (even positive voltage applied as long as it does not forward bias the body-source p-n junction) to allow for extremely fast switching. In the case of the DRAM sense amplifiers, the body was used to enhance and speed up the delicate operation of sensing cell data [71].

It would appear then that in most cases, FD devices would be preferred due to the enhanced performance and reduction of FBE and the kink effect over PD devices. Studies at the IBM Semiconductor Research and Development Center have shown that undepleted SOI results in better short channel effects than ultra-thin FD SOI and that FD SOI has very rigid contraints placed upon its device design [73]. The team used the FIELDAY device simulator to observe the FBE on a relatively thick 150nm PD SOI film compared to an ultra-thin 25nm FD device. FBE result because of a parasitic bipolar transistor that can be drawn between the source, body and drain in an SOI device. Varying the carrier lifetime in the simulator is the same as adjusting the bipolar gain. For the ultra-thin device, the subthreshold curves remain unchanged while the carrier lifetime was varied over five orders of magnitude, from $\tau e = \tau p = 10^{-4}$ to 10^{-9} s. However, if the SOI thickness is varied even slightly (25-50nm), the I_{off} becomes unacceptable. Reducing the carrier lifetime, this SOI thickness dependency was significantly reduced as shown in Figure 32.



(a) (b)

Figure 32. Id vs. Vgs characteristics for (a) carrier lifetime variation of an FD SOI device and (b) SOI thickness variation [73].

Thus, in order to keep the npn (for n-channel MOSFET in this case) parasitic effects down, the SOI thickness must be made "impractically thin". In general, bulk CMOS requires high substrate doping (10^{18} cm^{-3}) for elimination of SCE as mentioned previously. This results in a depletion region width of:

$$x_d = \left[\frac{2K_s\varepsilon_0(\Psi_0 + V_R)}{qN_d}\right]^{1/2} = 65nm \qquad \textbf{Equation (16)}$$

Thus, the SOI thickness must be less than 65 nm. The IBM team goes on to show that a PD device with retrograde doping in the channel and source and drain extension halos was used to significantly reduce the floating body effects to a tolerable level as carrier lifetime was varied over the five orders of magnitude (see Figure 33).

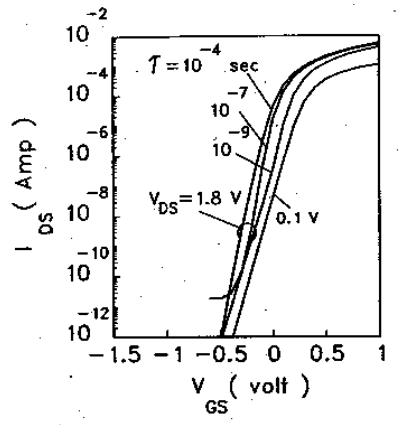


Figure 33. Id vs. Vgs dependency on carrier lifetime for PD SOI device with retrograde channel and halo implants [73].

c. Double Gate Structures

To further increase the performance of SOI structures, use of the substrate below the BOX was biased to create a ground plane below the silicon layer. This plane provided termination of the drain to source electric field which gives rise to hot carrier degradation in bulk CMOS. Later this substrate directly below the BOX was replaced with poly-silicon to form a second gate which was actively used during operation to create a second channel below the silicon. This second gate was utilized as early as 1987 [74]. H. S. Wong describes this as an evolution of the SOI technology as shown in Table VII [68].

Table VII. The evolution of SOI technology [68].

Variant	Strengths	Weaknesses
Partially depleted SOI	1. Channel design bulk-like 2. VT insensitive to BOX interface	1. Very susceptible to floating body effects (but solutions are available) 2. Same scaling constraints as bulk
Fully depleted SOI	1. Elimination of floating body effects 2. Elimination of punch-through currents 3. Elimination of drain-body tunneling	1. VT sensitive to SOI thickness and back interface 2. Back-channel potential may be influenced by drain voltage 3. Difficulty of contacting thin SOI
Ground Plane (GP)	1. Same as FD SOI 2. GP shields channel from drain 3. GP permits electrical control of VT 4. GP may be used as second gate	1. VT sensitive to SOI thickness 2. Difficulty of contacting thin SOI 3. Degradation of subthreshold slope by close GP
Double Gate (DG)	1. Maximum electrostatic control of channel and best scaling potential 2. Best current drive and performance 3. OR logic function within single device	1. Difficult to fabricate 2. Mis-aligned top and bottom gates result in extra capacitance and loss of current drive 3. VT control difficult by conventional means
Stacked SOI (ST)	1. High functional density 2. Shorter wires therefore higher performance and lower power	1. Fabrication complexity 2. Difficult to cool

The fabrication of the dual gate SOI transistor is made in one of three orientations as shown in Figure 34.

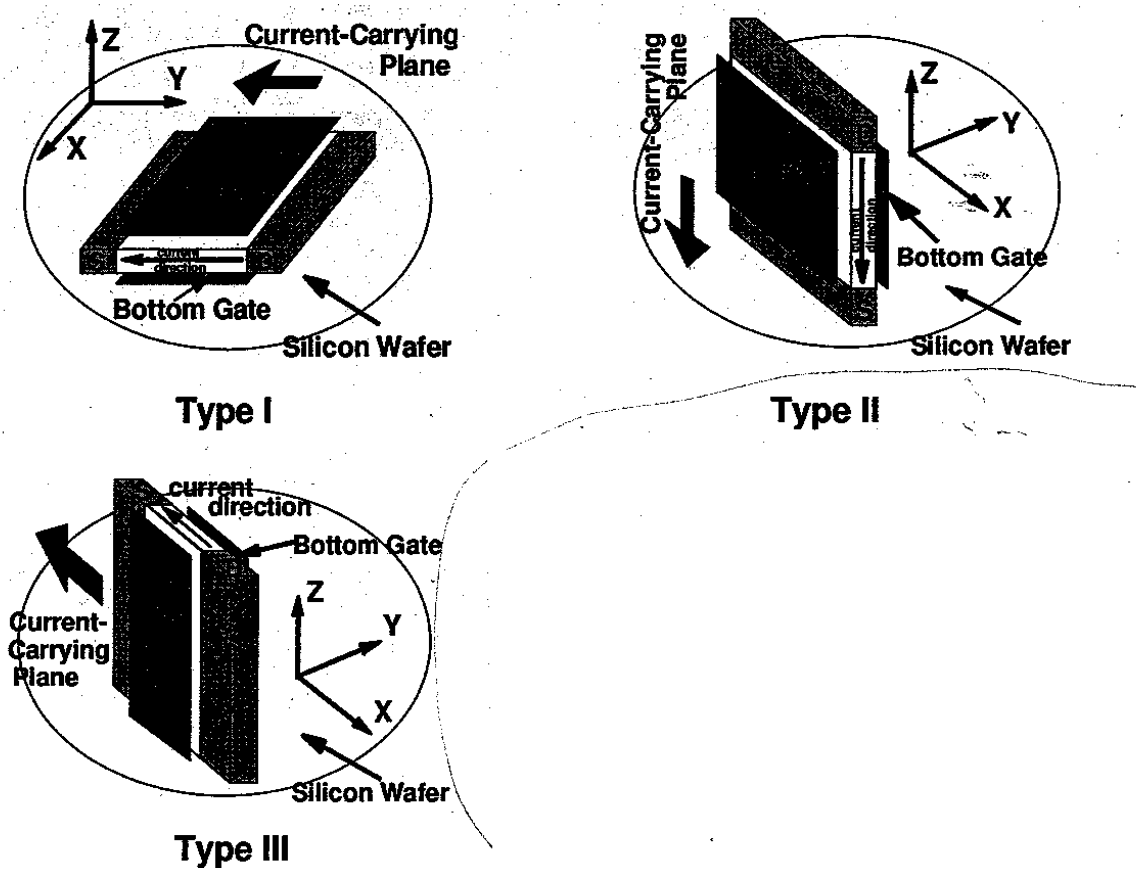


Figure 34. Different process topologies for DGFET [68]

Type I is of the traditional lateral transistor orientation with the gate parallel to the silicon [100] direction. The advantage of this fabrication technique is that the channel thickness is dependent upon the SOI process. Type II is a vertical transistor whose main advantage is that the chip density can be greatly increased but performance is secondary since lithography controls the channel thickness. In addition, it is difficult to make contact to the source node since it is buried in the silicon. The third type solves this problem by rotating the type II DGFET about the gate by 90 degrees [68].

d. DGFET Device Operation

In addition to the screening effect that the second gate provides for the drain to source electric field lines to eliminate SCE, actively controlling the bottom of the SOI layer as a second gate provides better circuit performance. An early study of the second gate structures claimed that the two gates used in tandem create "volume inversion" or that the entire SOI region between source and drain from first gate to second gate is strongly inverted [74]. This results in a channel which has a cross sectional area equivalent to the cross-sectional area of the SOI layer itself. The group used a device simulator and applied a potential Vg1 to the first gate and a potential Vg2 = KVg1 to the second gate in order to create the same channel characteristics at the back surface of the SOI as at the top surface. The factor K accounts for the difference in gate thickness of top and bottom oxides surrounding the SOI. If the SOI film thickness is greater than the depletion widths induced by the two gate voltages, then the increase in current driveability over a single gated SOI device of the same dimensions and bias point is a factor of 2. This accounts for the second channel on the bottom of the SOI layer:

$$I = \left(WV_D / L\right)\left(\mu_1 C_{ox1}\left(V_{g1} - V_{t1}\right) + \mu_2 C_{ox2}\left(KV_{g1} - V_{t2}\right)\right) \qquad \textbf{Equation (17)}$$

However, for thin SOI layers, the entire film can move into the strong inversion regime. This results in a tremendous improvement in current driveability for several reasons. Since the channel has expanded throughout the SOI layer, surface interactions such as scattering have been reduced. These increases in carrier mobility and transconductance result in a drain current improvement of 2.5 to 3 times. The improvement in the subthreshold Id-Vgs current slope due to the second gate is shown in Figure 35.

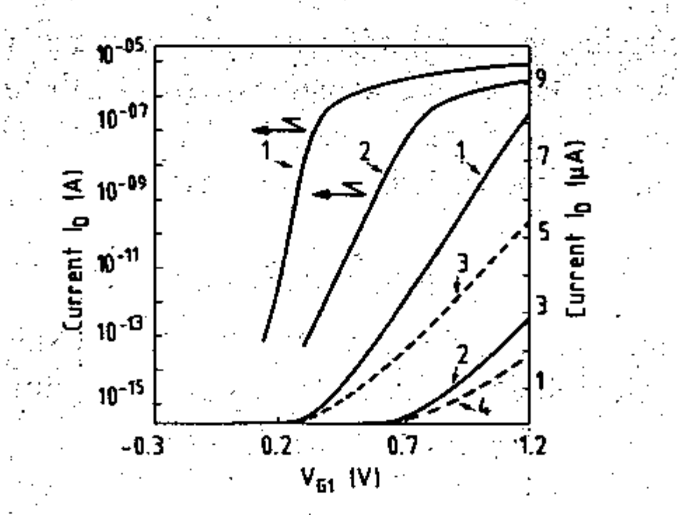


Figure 35. Improvement of Id vs. Vgs curves in subthreshold region (left vertical axis) and large Vgs region (right vertical axis) for DGFET (curve 1) over bulk CMOS (curve 2) [74].

2. Vertical Transistor

In the near future, the gate length will be less than 100 nm. The ability to make the gate length this short will be a challenge for optical lithography [75]. Optical lithography can produce gate sizes down to the sub 0.2um region. Lithography can make gate lengths in the range 100-120 nm using phase shifting techniques [76]. Other methods have been proven to work but do not seem attractive. For example E-beam lithography is a slow process but has produced gate sizes down to few a 10nm [75,76]. E-beam is mainly used in laboratories and for mask etching [76]. Also X-ray lithography has allowed gate lengths down to the 30nm region but is too expensive for mass production [75,76]. An intriguing alternative is the vertical transistor.

Shallow trench etching and epitaxy define the channel length of a vertical transistor. Atomic layer growth is used to deposit the epitaxial layers, which form the channel. The gate size is not dependent on lithography. The channel length would depend on the epitaxial layers [76].

Short channel lengths, however, are not the main advantage of the vertical transistor [75]. As planar transistors are scaled down, the width becomes smaller which causes the drive current to be less. Vertical transistors offer a larger width than planar transistors [77]. The increase in packing density is appealing to memory products such as DRAMs and ROMs. DRAMs will gain much higher density packing for the same gate dimensions as a conventional lateral transistor. With the use of vertical transistors in DRAM, a larger transistor could be used and still have the same size memory cell. Using a larger gate length will decrease short channel effects such as lower I_{off} [75]. The diagram in Figure 36 below shows a cross section schematic of an epitaxially grown channel vertical transistor.

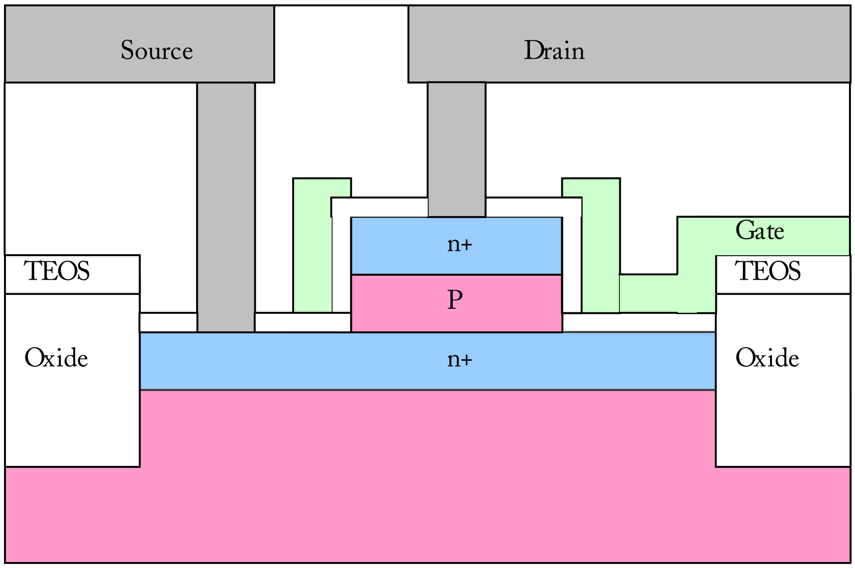


Figure 36. Schematic cross section of a n-type vertical transistor.

a. Reduction in Area

The layout below in Figure 37 shows a planar transistor compared to a vertical transistor. The gate is represented by the dashed lines. Both layouts have a single transistor with a width of 8F (F stands for feature size). The planar transistor has an area of 40 F^2 versus 16.5 F^2 for the vertical transistor. This is almost a 60% reduction in area.[75] Drive current is directly proportional to the width of a transistor. Both types of transistors have the same amount of drive current [77].

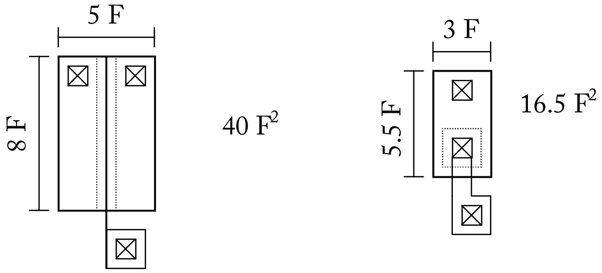


Figure 37. Comparison of the area of a planar transistor (left) verses vertical transistor (right).

Another example of the vertical transistor application is stacking two transistors to create the inverter shown below in Figure 38. The integration factor can be improved by a factor of 5 [75]. In the inverter the PFET has a greater width than the NFET. This should help the performance of the inverter. One of the transistors has to be wider to make contact to the middle section for Vout. Unfortunately this causes more area for the inverter. It is beneficial to make the PFET the lower transistor for more drive current since PFETs have a low mobility μ.

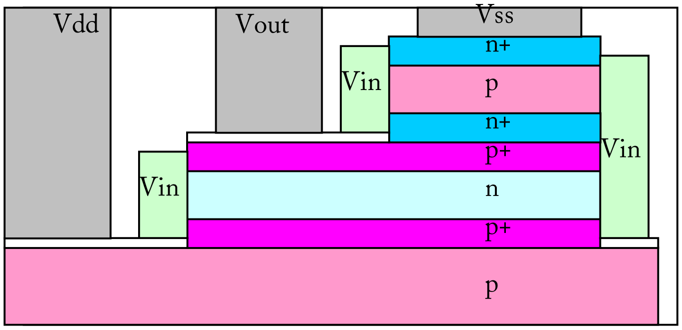


Figure 38. Cross-section of a vertical inverter.

b. Device Characterzation

Fabricated vertical transistors showed results similar to advanced planar transistors [78]. Reference [78] manufactured vertical transistors with a vertical gate length of 70 nm. The data showed high values for transconductance of 800 uS/um with voltage operation of 1.5 V. The maximum theoretical value is Cox*vsat=690uS/um. They claim the high value is due to ballistic overshoot effects. The threshold voltage for this transistor was 0.4 V which is comparable to a planar transistor. The subthreshold slope for a vertical transistor with a gate length of 170 nm was 125mV/dec, which is worse than a typical planar transistor. They believe this could be improved with process enhancement. They also noticed an increase in the slope of an Id vs. Vd chart at Vd=1.5 V. They attributed this upward kink to the floating channel region [78] The channel in this type of vertical transistor cannot have a bias; therefore the channel floats during operation.

c. Surrounded Gate Structure

Another kind of vertical transistor is the surrounded gate transistor. The difference between this type and the previous transistor is that the substrate is part of the channel. The channel is a pillar of silicon, and the channel is the same silicon as the substrate. The source is implanted on the side of the vertical gate in the substrate. The advantage of the gate being on the sides of the channel in a vertical direction is that the channel length can be increased by making the transistor taller.

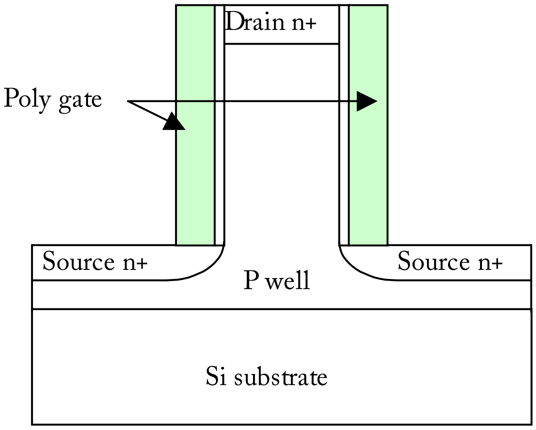


Figure 39. Surround gate structure.

This type of vertical transistor shows the subthreshold swing is better than lateral transistors. Subthreshold swing is defined as the amount of gate voltage decrease needed for one decade decrease in Ids. A slope of 72 mV/decade was measured [77] for a vertical transistor compared to a typical planar transistor of 98 mV/decade. The results showed that the subthreshold slope is dependent on the diameter of the vertical pillar that creates the channel [77]. To explain this decrease, a simple model is proposed [77].

The ideal model represents the vertical stack as a cylinder with the gate surrounding the channel stack. The depletion region is the striped area in Figure 40 and the depletion capacitance is given by Equation (18).

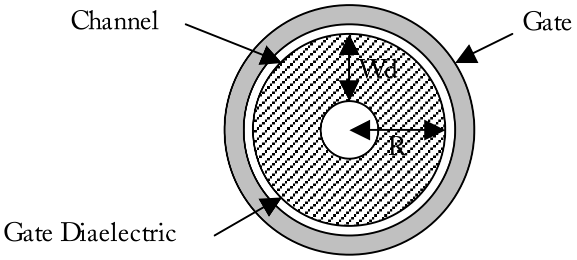


Figure 40. Pillar channel showing depletion length Wd.

$$C_d = -\frac{\partial Q_b}{\partial \phi_s} = \frac{\varepsilon_{si}}{R \cdot \ln \frac{R}{R - W_d}}$$ Equation (18)

The depletion capacitance is given for R≥W_d.

In Equation (18), Q_b represents the charge in the depletion region, R is the radius of the channel, W_d is the length of the depletion region, φ_s is the potential at the surface, and ε_s is the dielectric constant of the silicon. Equation (18) shows that C_d depends on the radius of the depletion region. Equation (18) is differentiated with respect to R which yields for R≥W_d

$$\frac{\partial C_d}{\partial R} = -\varepsilon_{si} \left(R \cdot \ln \frac{R}{R - W_d} \right)^{-2} \cdot \left(\ln \frac{R}{R - W_d} - \frac{R}{R - W_d} + 1 \right) \geq 0$$ Equation (19)

Equation (19) shows that as R goes to infinity, the capacitance approaches the capacitance of a planar transistor. The equation also shows that the capacitance is zero when R=W_d. The subthreshold swing can be approximated by the following equation where C_{ox} is the oxide capacitance,

$$S = \frac{kT}{q} \cdot \ln 10 \left(1 + \frac{C_d}{C_{ox}} \right)$$ Equation (20)

Equation (20) indicates small values of C_d will produce a small S. To maximize the performance of a vertical transistor, the depletion region should be of the same magnitude as the radius of the channel. A small S will help reduce I_{off} current because the slope is greater in the subthreshold region. Also the equation indicates that an increase in gate voltage will not deplete the channel anymore once Wd=R. The extra voltage will only increase the number of carries in the channel and increase the surface potential at the gate. The subthethreshold swing will reach an ideal value of 60 mV/decade when Wd=R since kT/q* ln 10 = 60. This is assuming no traps are in the gate oxide at the interface [77].

The surrounded gate structure shows very little substrate bias effects when the Wd=R compared to the planar transistor. The reason for this is because the channel pillar is completely depleted. The substrate bias only effects the bottom of the channel cylinder. When Wd<R, the substrate bias will have more of an effect on the channel. For high reliability circuits this effect will be a benefit since fluctuation in substrate bias will not effect the transistor [77].

d. Vertical Transistor for DRAM Application

DRAM has kept up the pace for packing more memory cells on a single chip. DRAM has been shrinking in area by a factor of four every three years. The way DRAM increases cell density is by shrinking the cell by 2/3 of the original cell. The traditional DRAM cell consists of a planar transistor and a storage cell. The storage cell is either a stack capacitor or a trench capacitor. The conventional DRAM cell has an optimal area of $8F^2$ per cell. The chart in Figure 41 shows the cell size verses generation of memory. The chart indicates that memory will not be able to use $8F^2$ cell size after 256 Meg. The optimal cell size for DRAM is a $4F^2$ cell size. The one way to achieve a $4F^2$ cell size would be from the use of a vertical transistor [79].

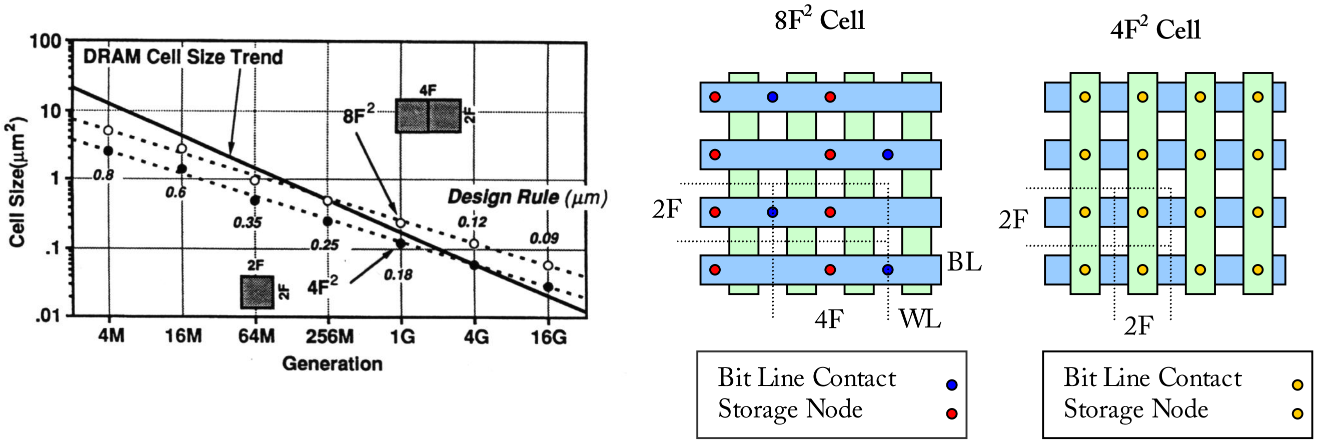


Figure 41. Dram cell size trend (left) and layout of a $8F^2$ and $4F^2$ cell [79].

The vertical transistor proposed for this $4F^2$ cell would be similar to the surrounded gate transistor mentioned above. The transistor would be surrounded by the word line. The capacitor storage would be directly above the vertical transistor. An advantage of using the surrounded channel as mentioned in the previous section is that the vertical transistor can have a low subthreshold voltage swing. This is important to have longer retention time. The channel conductance is expected to increase by volume inversion. The vertical transistor should reduce short channel effects. If the DRAM is made on SOI, this will reduce the requirement for storage capacitance since the cell should have less leakage and reduce parasitic capacitance.

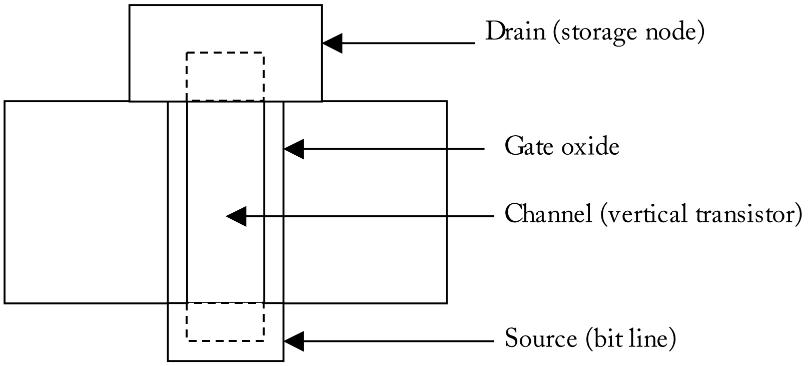


Figure 42. $4F^2$ cell cross-section.

The above proposed vertical $4F^2$ cell does have some barriers to overcome. The word line, bit line, and channel have a very low tolerance for misalignment. A self alignment process would have to be used [79]. Another problem with the vertical transistor would be the hole size for the channel itself, which has to be smaller than the word line, and also an insulator will have to be grown between the gate and the channel. The holes in the gate will increase the resistance of the word line. The $4F^2$ cell uses an open bit line scheme. In the folded bit line scheme, the bit lines are crossed for noise cancellation. Therefore, the open bit line scheme has more noise. A way to compensate this problem is to increase the read out voltage of the bit line. The read out voltage is defined as:

$$Vread-out=\frac{Vcc}{2\left(1+\frac{Cb}{Cs}\right)} \qquad \textbf{Equation (21)}$$

Vcc is the supply voltage, Cb is the bit line capcitance, and Cs is the storage capcitance. The easiest way to increase Vread-out is by not increasing Vcc (increase short channel effects) but to lower Cb. Cb can be decreased by using SOI. As the buried oxide increases, the capacitance between bit line and substrate decreases [79].

Another problem with $4F^2$ cell size is the size itself. In order to get the needed storage capacitance, the stack capacitor will need a towering height. A solution for this is to use a higher dielectric constant material such as BST [79]. Also using SOI will help to allow a lower storage capacitance [79].

3. Cu Metallization/Low K Dielectric Insulators

As the effective channel length of a MOSFET decreases, the carrier transit time across the length of the channel also decreases. This reduces the so-called gate delay and obviously leads to faster devices. In addition to this "intrinsic" delay, one must also consider the RC interconnect delay associated with the metal interconnects and insulating dielectrics found in all IC chips. The metal interconnects carry current to and from the active devices while the dielectrics electrically isolate interconnects from one another and provide mechanical stability. The RC interconnect delay can be approximated by the following equation if one ignores the effects of fringing capacitances [54]

$$RC=\frac{\rho}{t_m}\frac{L^2\varepsilon_{ILD}}{t_{ILD}} \qquad \textbf{Equation (22)}$$

where ρ is the interconnect resistivity, t_m is the interconnect thickness, L is the interconnect length, ε_{ILD} is the interlayer dielectric (ILD) permittivity and t_{ILD} is the interlayer dielectric thickness. For a given t_m and t_{ILD}, Equation (22) indicates that RC can only be reduced by decreasing ρ, L and/or ε_{ILD} [54].

It is widely accepted that at very small dimensions, the total circuit delay (intrinsic gate delay + RC interconnect delay) is dominated by the RC interconnect delay [7,54,80]. According to Figure 43, which shows the intrinsic gate delay and RC interconnect delay versus generation (i.e., channel length), the interconnect delay dominates the gate delay for Al-based interconnects and SiO_2 dielectrics as the channel length approaches 100 nm [7]. The total delay can be reduced by switching to metals with lower resistivity, such as Cu, and dielectrics with a lower dielectric constant. This concept is illustrated in Figure 43, where a decrease in the interconnect delay and improvement in the overall delay are expected for Cu interconnects and low dielectric constant (K) insulators.

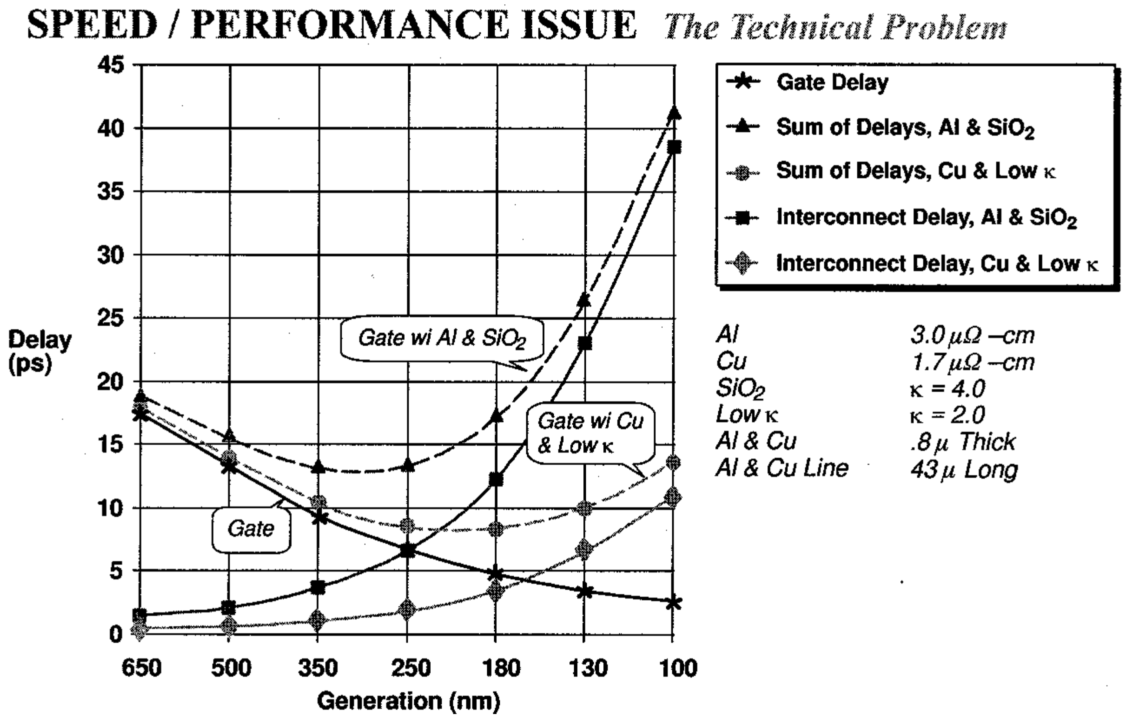


Figure 43. Calculated gate and interconnect delay versus technology generation [7].

As the channel length approaches 100 nm, chip performance will also be limited by power consumption. Power dissipation in interconnects is primarily driven by the total wiring capacitance, which contributes to the dynamic power dissipation [81]. For example, the dynamic power dissipation, P_D, of a simple CMOS inverter loaded by a capacitor C is equal to fCV_{DD}^2. Here, f is the switching frequency of the inverter and C represents the sum of the internal device capacitances, the capacitance of the interconnect wire between the inverter output and the input of other logic gates, and the total input capacitance of these other logic gates [82]. Therefore, while a change in the interconnect resistivity only directly affects the RC interconnect delay, a change in the interlayer dielectric constant directly impacts both the RC interconnect delay and the power consumption [81].

The transition from Al interconnects and SiO_2 insulators to Cu interconnects and low K dielectric constant insulators will vary from company to company. For example, while IBM and Motorola are producing Cu interconnect technologies before employing low K dielectric materials, Intel and Texas Instruments have chosen to work with Al interconnect technologies and implement low K dielectric materials [83]. It is estimated that most companies will incorporate Cu interconnects and low K dielectric materials in IC chips when feature sizes reach 130 nm in 2003 [83].

In this section, we explore the advantages and disadvantages of using Cu interconnects and low K dielectric constant insulators in IC chips. We also discuss the challenges that face the semiconductor industry as it attempts to incorporate these new materials into mass production.

a. Properties of Cu

The property of Cu that makes it most attractive as a replacement for Al in IC chips is undoubtedly its lower resistivity. Nevertheless, the choice of which metal should be used for future interconnect wiring is not only determined by the resistivity of the material. In addition, other material properties, such as Young's modulus and thermal conductivity, must be taken into account. Table VIII compares the properties of several metals that have been considered for interconnect applications.

Table VIII. Comparison of properties of possible interlayer metals [54].

	Metal				
Property	**Cu**	**Ag**	**Au**	**Al**	**W**
Resistivity (μΩ-cm)	1.67	1.59	2.35	2.66	5.65
Young's Modulus x 10^{-11} dyn/cm^2	12.98	8.27	7.85	7.06	41.1
TCR x 10^3/K	4.3	4.1	4	4.5	4.8
Thermal Conductivity (W/cm)	3.98	4.25	3.15	2.38	1.74
CTE x 10^6/°C	17	19.1	14.2	23.5	4.5
Melting Point (°C)	1085	962	1064	660	3387
Specific Heat Capacity (J/kg K)	386	234	132	917	138
Corrosion in Air	Poor	Poor	Excellent	Good	Good
Adhesion to SiO_2	Poor	Poor	Poor	Good	Poor
Delay (ps/mm)	2.3	2.2	3.2	3.7	7.8
Thermal Stress per Degree for Films On Si (10^7 dyn/cm^2-°C)	2.5	1.9	1.2	2.1	0.8
Note: Delay = RC = 34.5 R_s (ps/mm) for 1 mm-length conductor on 1 μm thick SiO_2.					

Although Cu will eventually replace Al in all high performance IC chips, integration of Cu into VLSI and ULSI circuits poses serious challenges to the semiconductor industry. Cu diffuses readily through SiO_2 and other dielectric materials, which can cause an increase in junction leakage and reduction in gate oxide yield and reliability if it diffuses into active device regions [84]. Also, Cu does not adhere well to SiO_2 or other dielectric materials, and does not possess a passivating oxide, such as Al_2O_3 in the case of Al [84,85]. The absence of a passivating oxide increases the susceptibility of Cu to corrosion. Finally, Cu lacks a volatile compound at room temperature, which makes conventional reactive ion etch (RIE) patterning techniques unsuitable for this material [54,84].

b. Cu Electromigration

Although it is not listed explicitly in Table VIII, the electromigration resistance of a metal is of utmost importance in determining the reliability of interconnects. The electromigration resistance of Cu is believed to be far superior (10-100x higher) to that of Al [54,55]. An improvement in the electromigration lifetime leads to an increase in the allowed current density (i.e., more design flexibility) or to a reduced interconnect line width (i.e., improved wiring density).

As mentioned in Section III.C.3., electromigration is electronic current induced atomic diffusion. At sufficiently high current densities, momentum is transferred from the conducting electrons to the metal ions. The atomic flux, J, that results from this momentum transfer is given by the following equation [86,87]

$$J = \frac{nj\rho eZ^{*}D}{kT} \qquad \textbf{Equation (23)}$$

where n is the atomic density, j is the current density, ρ is the conductor resistivity, e is the electronic charge, Z^* is the effective charge number, D is the atomic diffusivity, k is Boltzmann's constant and T is the temperature. According to Equation (23), the atomic flux is directly proportional to Z^*, ρ and D. Generally speaking, the higher the melting point of a particular metal, the lower is the diffusivity. Z^* is a measure of the amount of interaction between the conducting electrons and the metal ions. A higher degree of interacton results in a larger atomic flux. For Bulk materials at 100 °C [54], the product $Z^*\rho D$ (in μΩ-cm^3/s) is found to equal 2.84-7.07 x 10^{-25} for Ag, 3.62-9.12 x 10^{-19} for Al, 3.05-3.83 x 10^{-26} for Au, and 1.3-1.5 x 10^{-29} for Cu. Therefore, at least in principle, Cu appears to be quite attractive from an electromigration standpoint.

It is important to realize, however, that the electromigration resistance of thin films is not only dependent on bulk properties. Factors that play an important role in Al electromigration, such as interfacial diffusion [88], crystallographic texture, grain size, and impurities [52], may also effect the electromigration performance of Cu interconnects. For example, while tin (Sn) and zirconium (Zr) impurities in Cu drastically increase the electromigration lifetime, magnesium (Mg) degrades the electromigration performance of Cu [89]. Also, Cu is known to adhere poorly to SiO_2 (see Table VIII), which may cause interfacial diffusion to be the primary atomic transport mechanism [85]. This may lead to lower than expected electromigration lifetimes since interfacial diffusion is considered to be a much faster diffusion pathway than lattice or bulk diffusion.

c. Heating Effects in Cu

Joule heating of interconnects is the temperature rise caused by the passage of current. Joule heating occurs because the metal lines heat up faster than the heat is dissipated to the surroundings (i.e., dielectrics and/or Si substrate). Qualitatively speaking, Joule heating increases as the current density increases while less Joule heating occurs in low resistivity or high thermal conductivity materials. Therefore, as suggested by Table VIII, the heat dissipated in Cu interconnects (for a given current density) is expected to be less than that in Al interconnects. This should lead to improvements in temperature accelerated reliability failure mechanisms, such as electromigration, as wiring dimensions are scaled down and current densities are increased.

d. Cu Integration

The standard process for depositing and patterning Al interconnects is to deposit the Al over the entire wafer and then RIE etch regions not protected by a photoresist mask [54]. As mentioned above, this approach is not suitable for Cu because it does not possess a volatile compound at room temperature. In the case of Al etching, aluminum tri-chloride ($AlCl_3$) is the volatile compound formed between the RIE gas (where Cl_2 + HCl or Cl_2 + BCl_3 are the most common etch chemistries) and the Al interconnects. Therefore, efforts have turned to single or dual damascene techniques for integrating Cu metallization into VLSI and ULSI circuits. In the damascene process, trenches or openings are etched into the underlying dielectric and then filled with metal. Single damascene involves filling interconnect trenches and interlevel via openings in separate steps, while dual damascene allows both to be filled simultaneously [90]. Planarized top surfaces are achieved by removing undesired Cu material using chemical mechanical polishing (CMP). Figure 44 shows a schematic cross-sectional view of planarized damascene Cu interconnects. The first level metal shown in Figure 44 is obtained by a single damascene process while the second and third level metals are obtained by a dual damascene process.

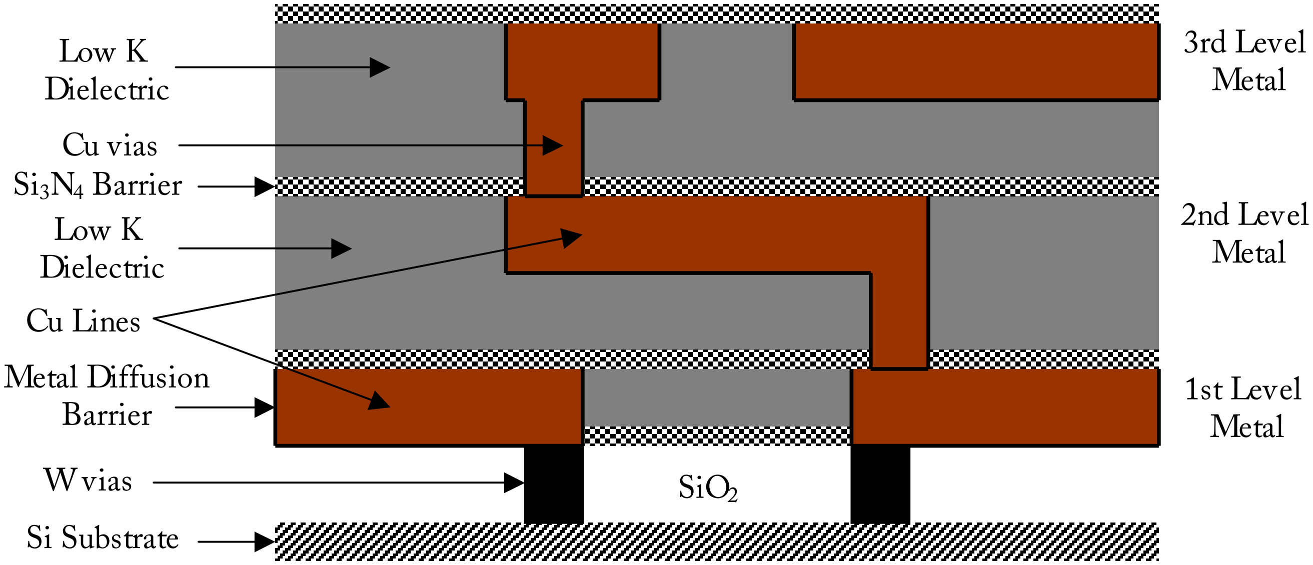


Figure 44. Schematic cross section of damascene Cu metallization with low K dielectric insulator.

Techniques for depositing Cu involve physical vapor deposition (PVD), chemical vapor deposition (CVD), or electroplating or electroless plating [84]. PVD methods, such as evaporation and sputtering, are less than ideal for damascene integration since it is difficult to fill high aspect ratio features [54]. High aspect ratio features have become more common in state of the art IC chips as interconnect dimensions are reduced. CVD methods, on the other hand, overcome the filling problems of PVD techniques since Cu is deposited in the gaseous phase [91]. Recently, electroplated Cu has been reported for sub-0.25 μm technologies [55,92]. A significant advantage of plating over PVD or CVD processes is related to its lower thermal budget. While Cu PVD and CVD deposition temperatures are approximately 350-450 °C and 250 °C, respectively, Cu plating is accomplished at temperatures < 100 °C [93]. Electroplating of Cu is achieved by submersing the Si wafer into a liquid bath containing cupric ions (Cu^{2+}), and allowing a current to flow from a contact at the wafer edge to every surface on the wafer that is to receive deposited material [94]. Prior to electroplating, it is necessary to cover each of these surfaces with a Cu seed layer that serves to conduct the applied current during the plating process [94]. The seed layer is typically very thin (10-20 nm) and is deposited by either PVD or CVD techniques.

Since Cu does not adhere well to SiO_2 or other dielectrics, and diffuses readily through these materials, Cu integration employs adhesion promoters and diffusion barriers. Adhesion promoters cause chemical bonding between the material being deposited and the substrate surface [54]. Diffusion barriers, such as refractory metals, are deposited into the trench and via openings prior to Cu deposition (see Figure 44). Diffusion through refractory metals, such as tantalum (Ta) and tungsten (W), is poor due to their high melting temperatures (i.e., strong interatomic bonding). A major drawback of using these barrier layers is the increase in the interconnect resistance. The resistivity of barrier layers is typically much greater than that of Cu. Therefore, in order to maintain the advantage of the resistivity of Cu over Al, it is necessary to limit the barrier thickness to less than 20 nm [54]. As dimensions are scaled down in future CMOS technologies, it will be a significant challenge for the semiconductor industry to ensure that a thin and continuous barrier layer exists along the trench and via surfaces. Finally, following Cu CMP, a dielectric barrier, such as silicon nitride (Si_3N_4), is deposited over the entire wafer in order to prevent Cu from diffusing out of the top metal surface and into the surrounding insulator (see Figure 44).

e. Types of low K Dielectric Materials

The major disadvantage of conventional SiO_2 as an interlayer dielectric material is its relatively high dielectric constant (K≈4). Since the RC interconnect delay and power consumption are directly related to the interlayer dielectric constant, there has been much effort in recent years to develop low K dielectric insulators. Table IX classifies various low K dielectric materials.

Table IX. Classification of Low K Dielectric Materials [81].

Material Class	Dielectric Constant	Deposition Type
Inorganic	Range 2.8-5.0	
CVD SiO_2	3.9-5.0	CVD
Thermal SiO_2	3.9	Oxidation
Modified SiO_2 (e.g. fluorinated)	2.8-3.9	CVD/ECR
Si (BN)	>2.9	CVD
Organic	Range 1.3-3.9	
Polyimides	2.9-3.9	SOG/CVD
Fluorinated Polyimide	2.3-2.8	SOG/CVD
Fluoropolymers	1.8-2.2	SOG
Aerogels (microporous)	1.3-2.2	SOG
Air Bridge	Range 1.0-1.2	

The first real efforts to reduce the interlayer dielectric constant involved fluorine-doped SiO_2 films. Fluorinated silicate glass, or FSG, films contain fluorine in small concentrations. A minimum value of $K \approx 3$ is found for 10 atomic percent (at %) fluorine in SiO_2 [95].

Although organic dielectric materials can achieve a dielectric constant < 2 (see Table IX), many of these films suffer from thermal stability problems for temperatures above 350 °C, and therefore may not be suitable for IC applications. Thermal stability is related to the glass transition temperature (T_g), which is the temperature above which a material changes from the solid to the glassy state [95]. Most organic films have T_g< 450 °C, which may present serious problems for integration since some IC processing steps exceed this temperature [95]. One organic film that shows potential for IC use is Dow Chemical's SiLK. A spin-on polymer with K=2.65 and thermal stability at temperatures > 450 °C, SiLK is actively being pursued by many companies for future CMOS technologies [83]. Recently, attempts have been made to incorporate air (K=1) into high dielectric constant materials to lower the overall dielectric constant. In Aerogels, Xerogels and foams, air is trapped as bubbles in a gel or polymer [95]. Porous materials such as these possess a very low dielectric constant (K as low as 1.3) as well as the rigidity and thermal stability requirements for IC implementation [81,95,96].

f. Heating Effects in Low K Dielectric Systems

Heat generated by interconnects (i.e., Joule heating) is conducted through the surrounding dielectric to the Si substrate. The higher is the thermal conductivity of the encapsulating insulator, the easier it is for heat to be removed from current-carrying interconnects. A major consequence of lowering the interlayer dielectric constant is that the thermal conductivity is also lowered [97]. For example, the thermal conductivity of CVD SiO_2 is 1.2 W/m°K while that of polyimide is only 0.24 W/m°K [98]. Therefore, as interconnect dimensions are scaled down and current densities increase, Joule heating effects are likely to become a serious concern in low K dielectric systems.

g. Integration of Low K Dielectric Materials

Figure 44 shows a schematic cross-sectional view of damascene Cu metallization combined with a low K dielectric constant insulator. Integration of low K dielectric constant materials into state of the art CMOS technologies is a tremendous challenge for the semiconductor industry, which, for the last thirty years, has been content with implementing SiO_2 interlayer dielectrics. There are many requirements for dielectric materials that must be considered in order to ensure chip performance and reliability. These requirements include the following [81]: electrical (low dielectric constant, low leakage current, high breakdown field), mechanical (ability to withstand large stresses, high crack resistance), chemical (low moisture absorption, high etch selectivity, good metal adhesion). Inevitably, most materials considered for interlayer dielectric applications meet some but not all of these requirements. For example, while Aerogels exhibit a dielectric constant as low as 1.3, their chemical etch selectivity is poor due to their porous structure [81]. Etch selectivity is extremely important for dual damascene integration, since trenches are etched into the underlying dielectric before being filled with metal.

Deposition of dielectric films varies from material to material. As indicated in Table IX, most SiO_2 films are deposited by CVD methods, while most organic films are deposited by spin coating. A dielectric deposited by spin coating is referred to as spin on glass (SOG). In the spin coating process, a pre-polymer in solution is spun onto the Si wafer surface, followed by a curing cycle that changes the pre-polymer to a polymer [95]. SiLK is an example of an organic spin-on polymer. CVD organic dielectrics are also being developed, such as parylene-N ($K \approx 2.6$), parylene-F ($K \approx 2.2$-2.3), Teflon-AF ($K=1.93$), polynaphthalene-N ($K=2.4$), polynaphthalene-F ($K=2.3$), fluorinated amorphous carbon ($K<3$) and fluorinated hydrocarbon ($K=2.0$-2.4) [95].

IV. DISCUSSION AND CONCLUSIONS

Over the last thirty years, the semiconductor industry has produced Si ICs with progressively higher circuit speed and density. The ability of the microelectronics world to develop and manufacture ICs with smaller features is the main reason for increased chip performance and packing density. The minimum feature size for CMOS technology has decreased from 1 μm in the late 1980's to 0.18 μm this year, and is projected to reach 0.10 μm before the year 2005.

In going from one CMOS generation to the next, it is necessary to scale MOSFETs to smaller dimensions. There are many scaling approaches, such as constant electric field, constant voltage, constant electrostatic, subthreshold and off current scaling. All of the methods attempt to produce long channel behavior in a short channel device, and most are compromises between reality and ideal scaling. Depending on the application, one approach may be more appropriate than another. For example, off current scaling is useful for DRAM technologies since the retention time is very dependent on I_{off}. In general, combinations of each method may be utilized.

Short channel effects in sub-micron technologies, such as DIBL, punch through and mobility degradation, pose serious challenges for future MOSFET scaling. One of the most obvious consequences of scaling is the decrease in the threshold voltage as the channel length is reduced (i.e., DIBL). For digital applications, the threshold voltage must be at least 0.4 V in order to ensure acceptable off current and noise margin. Short channel effects are controlled in state-of-the-art MOSFET devices by employing source/drain and channel engineered structures. Features such as source/drain extensions, halo implants and retrograde well profiles have allowed MOSFETs to be scaled to dimensions that would have been unattainable with conventional device engineering. It is uncertain, however, whether these techniques will provide the same benefits for MOSFETs with channel lengths less than or equal to 0.1 μm. Therefore, new processes and device designs may be required to allow scaling into the deep sub-micron regime.

As the 0.1 μm regime approaches, many companies in the semiconductor industry are developing unconventional technologies that represent alternatives to traditional scaling approaches. SOI technology gives an increase in performance by reducing the parasitic source and drain capacitances. For the same channel length and gate oxide thickness, SOI results in a 15-50% increase in speed over traditional CMOS. This means the MOSFET performance does not have to be improved by reducing the gate oxide thickness. The vertical transistor represents a novel means of overcoming lithography limitations for $L_{eff} \leq 0.1$ μm. In addition to gate lengths not being dependent on lithography, vertical transistors offer larger width (i.e., increased drive current) and higher packing density as compared to conventional lateral transistors. In order to decrease the interconnect delay and improve the overall delay in IC chips, Cu metallization and low K dielectric materials will gradually replace Al metallization and SiO_2 insulators. Switching from Cu to Al interconnects directly improves the RC interconnect delay, while replacing SiO_2 with a low K dielectric constant material directly improves both the RC interconnect delay and the power consumption.

Predicting the limits of Si MOSFET technology has not proved very reliable over the years. For example, published studies in 1979, 1984 and 1988 placed the limit of scaling for channel length/ oxide thickness/V_{CC} at 0.5 µm/90 Å/1.5 V, 0.2 µm/100 Å/2.2 V, and 0.4 µm/110 Å/2.8 V [1]. These numbers are significantly larger than what is used today in state-of-the-art CMOS technology. The main reason for this has been a lack of knowledge about the physics at reduced dimensions. Extrapolation of data collected for channel lengths between 0.3-0.5 µm may not be adequate to predict the behavior of devices with channel lengths < 0.1 µm. It is clear that Si has limits like any other material, and the reliability of the gate oxide may eventually limit MOSFET scaling. But until the physics of device operation and reliability failure mechanisms (i.e., electron-electron scattering in hot carrier degradation) at very small dimensions are understood, predicting the limits of MOSFET scaling will continue to be inaccurate.

V. ACKNOWLEDGEMENTS

We would like to thank Dr. M. S. Shur, Dr. G. LaRosa, Dr. R. A. Wachnik, Dr. P. C. Wang, Dr. S. Saroop and Dr. G. A. Biery for reviewing this project.

VI. REFERENCES

1. C. Hu, Ultra-Large-Scale Integration Device Scaling and Reliability, J. Vac. Sci. Technol. B., Vol. *12*, 1994, pp. 3237-3241.
2. H. Iwai, Silicon MOSFET Scaling Beyond 0.1 Micron, Proceedings of the 21st International Conference on Microelectronics, Vol. **1**, 1997, pp. 11-18.
3. Y. Taur, et al., CMOS Scaling into the Nanometer Regime, Proceedings of the IEEE, Vol. **85**, 1997, pp. 486-504.
4. S. Thompson, P. Packan, M. Bohr, MOS Scaling: Transistor Challenges for the 21st Century, Intel Technology Journal, Q398, pp. 1-19.
5. J. B. Roldan, F. Gamiz, J. A. Lopez-Villanueva, P. Cartujo, J. E. Carceller, A Model for the Drain Current of Deep Submicrometer MOSFETs Including Electron-Velocity Overshoot, IEEE Transactions on Electron Devices, Vol. **ED-45**, 1998, pp. 2249-2251.
6. C. Hu, Future CMOS Scaling and Reliability, Proceedings of the IEEE, Vol. **81**, 1993, pp. 682-689.
7. The National Technology Roadmap for Semiconductors, 1997 Edition, Semiconductor Industry Association.
8. R. Keyes, "The Physics of VLSI Systems", Addison-Wesley, Wokingham, England, 1987.
9. R. H. Dennard, F. H. Gaensslen, H-N. Yu, V. L. Rideout, E. Bassous, A. R. LeBlanc, Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions, IEEE Journal of Solid-State Circuits, Vol. **SC-9**, 1974, pp.256-268.
10. S. Wolf, "Silicon Processing for the VLSI Era – Volume III – The Submicron MOSFET", Lattice Press, California, 1995.
11. G. Baccarini, M. R. Wordeman, R. H. Dennard, Generalized Scaling Theory and Its Appplication to a ¼ Micrometer MOSFET Design, IEEE Transactions on Electron Devices, Vol. **ED-31**, 1984, pp. 452-462.
12. J. R. Brews, et al., Generalized Guide to MOSFET Miniaturization, IEEE Electron Dev. Letts., Vol. **EDL-1**, 1980, p. 2.
13. J. R. Brews, K. K. Ng, R. K. Watts, "The Submicron Silicon MOSFET", Ed. R. K. Watts, Wiley Interscience, New York, 1989, Chapter 1.
14. S.M. Sze, "Physics of Semiconductor Devices", 2nd Ed., Wiley Interscience, New York, 1981.
15. S. M. Kang, Y. Leblebici, "CMOS Digital Integrated Circuits Analysis and Design", The McGraw-Hill Co., Inc., New York, 1996.
16. B. Yu, C. H. J. Wann, E. D. Nowak, K. Noda, C. Hu, Short Channel Effect Improved by Lateral Channel-Engineering in Deep-Submicronmeter MOSFET's, IEEE Transactions on Electron Devices, Vol. **ED-44**, 1997, pp. 627-634.
17. B. Davari, R. H. Dennard, G. G. Shahidi, CMOS Scaling for High Performance and Low Power – The Next Ten Years, Proceedings of the IEEE, Vol. **83**, 1995, pp. 595-606.
18. Y. Taur, Y.-J. Mii, D. J. Frank, H.-S. Wong, D. A. Buchanan, S. J. Wind, S. A. Rishton, G. A. Sai-Halasz, E. J. Nowak, CMOS Scaling into the 21st Century: 0.1 μm and Beyond, IBM J. Res. Develop., vol. **39**, 1995, pp. 245-259.
19. J. P. Uyemura, "Circuit Design For CMOS VLSI", Kluwer Academic Publishers, Boston, 1992.
20. B. G. Streetman, "Solid State Electronic Devices", Prentice Hall, New Jersey, 1990.

21. T. A. Fjeldly, M. Shur, Threshold Voltage Modeling and the Subthreshold Regime of Operation of Short-Channel MOSFET's, IEEE Transactions on Electron Devices, Vol. **ED-40**, 1993, pp. 137-145.
22. J. E. Chung, M-C. Jeng, J. E. Moon, P-K. Ko, C. Hu, Performance and Reliability Design Issues for Deep-Submicrometer MOSFET's, IEEE Transactions on Electron Devices, Vol. **ED-38**, 1991, pp. 545-554.
23. S. Matsuda, T. Sato, H. Yoshimura, Y. Takegawa, A, Sudo, I. Mizushima, Y. Tsunashima, Y. Toyoshima, Novel Corner Rounding Process for Shallow Trench Isolation Utilizing MSTS (Micro-Structure Transformation of Silicon), International Electron Device Meeting, Technical Digest, 1998, pp. 137-140.
24. W. Lee, T. Osakama, K. Asada, and T. Sugano, Design Methodology and Size Limitations of Submicrometer MOSFET's for DRAM Application, IEEE Transactions on Electron Devices, Vol. **ED-35**, 1988, pp. 1876-1884.
25. C. Fiegna, H. Iwai, T. Wada, M. Saito, E. Sangiorgi, D. Ricco, Scaling the MOS Transistor Below 0.1 μm: Methodology, Device Structures, and Technology Requirements, IEEE Transactions on Electron Devices, Vol. **ED-41**, 1994, pp.941-949.
26. M. J. Hargrove, S. Voldman, R. Gauthier, J. Brown, K. Duncan, W. Craig, Latchup in CMOS Technology, 1998 IEEE International Reliability Physics Symposium, Proceedings, pp. 269-278.
27. Y. Taur, T Ning, FEOL Technology Trend, Materials Chemistry and Physics, Vol. **52**, 1998, pp. 191-199.
28. Y. Taur, S. Wind, Y. Mii, Y. Lii, D. Moy, K. Jenkins, C. L. Chen, P. J. Coane, D. Klaus, J. Bucchingnano, R. RosenField, M. Thomson, M. Polcari, High Performance 0.1 um CMOS Devices with 1.5-V Power Supply, International Electron Device Meeting, Technical Digest, 1993, pp. 127-130.
29. C. Hu, S. C. Tam, F-C. Hsu, P-K. Ko, T-Y Chan, K. W. Terrill, Hot-Electron-Induced MOSFET Degradation - Model, Monitor, and Improvement, IEEE Transactions on Electron Devices, Vol. **ED-32**, 1985, pp. 375-385.
30. E. S. Yang, "Microelectronic Devices", McGraw-Hill, New York, 1988, pp. 285-294.
31. M. Shur, "Introduction to Electronic Devices", John Wiley & Sons, New York, 1996, p. 375.
32. F. C. Hsu, S. Tam, Relationship between MOSFET Degradation and Hot-Electron Induced Interface-State Generation, IEEE Electron Device Lett., Vol. **EDL–5**, 1984, pp. 50-52.
33. R. Woltjer, A. Hamada, E. Takeda, Time Dependence of p-MOSFET Hot-Carrier Degradation Measured and Interpreted Consistently Over Ten Orders of Magnitude, IEEE Transactions on Electron Devices, Vol. **ED-40**, 1993, pp. 392-401.
34. S. Ogura, P. J. Tsang, W. W. Walker, D. L. Critchlow, J. F. Shepard, Design and Characterization of the Lightly Doped Drain-Source (LDD) Insulated Gate Field-Effect Transistor, IEEE Transactions on Electron Devices, Vol. **ED-27**, 1980, pp. 1359-1367.
35. P. J. Tsang, S. Ogura, W. W. Walker, J. F. Shepard, D. L. Critchlow, Fabrication of High-Performance LDDFET's with Oxide Sidewall-Spacer Technology, IEEE Transactions on Electron Devices, Vol. **ED-29**, 1982, pp. 590-596.
36. J. Frey, Where do Hot Electrons Come From?, IEEE Circuits and Devices Magazine, Nov. 1991, pp. 31-37.

37. P. A. Childs, C.C.C. Leung, New Mechanisms of Hot Carrier Generation in Very Short Channel MOSFETs, Electronics Letters, Vol. **31**, January 1995, pp. 139-141.

38. M. Fischetti, S. Laux, Monte Carlo Study of Sub-Band-Gap Impact Ionization in Small Silicon Field–Effect Transistors, International Electron Device Meeting, Technical Digest, 1995, pp. 305-308.

39. J. Bude, T. Iizuka, Y. Kamakura, Determination of Threshold Energy for Hot Electron Interface State Generation, International Electron Device Meeting, Technical Digest, 1996, pp. 865-868.

40. J. J. Ellis-Monaghan, R. B. Hulfachor, K. W. Kim, M. A. Littlejohn, Ensamble Monte Carlo Study of Interface-State Generation in Low-Voltage Scaled Silicon MOS Devices, IEEE Transactions on Electron Devices, Vol. **ED-43**, 1996, pp.1123-1132.

41. S. E. Rauch, III, F. J. Guarin, G. LaRosa, Impact of E-E Scattering to the Hot Carrier Degradation of Deep Submicron NMOSFET's, IEEE Electron Device Letters, Vol. **EDL-19**, 1998, pp. 463-465.

42. Y. Kamakura, H. Mizuno, M. Yamaji, M. Morifuji, K. Taniguchi, C. Hamaguchi, Impact Ionization Model for Full Band Monte Carlo Simulation, J. Appl. Phys., Vol. **75**, 1994, pp. 3500-3506.

43. J. D. Bude, Gate Current by Impact Ionization Feedback in Sub-Micron MOSFET Technologies, Symposium on VLSI Technology, Technical Digest, 1995, pp. 101-102.

44. M. Shatzkes, M. Av-Ron, R. A. Gdula, Defect-Related Breakdown and Conduction in SiO_2, IBM J. Res. Develop., Vol. **24**, 1980, pp.469-479.

45. E. Rosenbaum, Oxide Reliability, 1996 IEEE International Reliability Physics Symposium, Tutorial Notes, Topic 6a, pp. 6a.1-6a.27.

46. R. Degraeve, Oxide Reliability, 1997 IEEE International Reliability Physics Symposium, Tutorial Notes, Topic 7, pp. 7.1-7.71.

47. D. J. DiMaria, E. Cartier, Mechanism for Stress-Induced Leakage Currents in Thin Silicon Dioxide Films, J. Appl. Phys. Vol. **78**, 1995, pp. 3883-3894.

48. C. Hu, Gate Oxide Scaling Limits and Projection, International Electron Device Meeting, Technical Digest, 1996, pp. 319-322.

49. J. H. Stathis, D. J. DiMaria, Reliability Projection for Ultra-Thin Oxides at Low Voltage, International Electron Device Meeting, Technical Digest, 1998, pp. 167-170.

50. J. R. Black, "Electromigration Failure Modes in Aluminum Metallization for Semiconductor Devices", Proceedings of the IEEE, vol. **57**, 1969, pp. 1587-1594.

51. C.-K. Hu, M. B. Small, P. S. Ho, Electromigration in Al(Cu) Two-Level Structures: Effect of Cu and Kinetics of Damage Formation, J. Appl. Phys., Vol. **74**, 1993, pp. 969-978.

52. I. Ames, F. M. d'Heurle, R. E. Horstmann, Reduction of Electromigration in Aluminum Films by Copper Doping, IBM J. Res. Develop., Vol. **14**, 1970, pp. 461-463.

53. C.-K. Hu, Electromigration Failure Mechanisms in Bamboo-Grained Al(Cu) Interconnections, Thin Solid Films, Vol. **260**, 1995, pp. 124-134.

54. S. P. Murarka, S. W. Hymes, Copper Metallization for ULSI and Beyond, Crit. Rev. Solid State Mater. Sci., Vol **20**, 1995, pp. 87-124.

55. D. Edelstein, et al., Full Copper Wiring in a Sub-0.25 µm CMOS ULSI Technology, International Electron Device Meeting, Technical Digest, 1997, pp. 773-776.

56. S. C. Williams, R. B. Hulfachor, K. W. Kim, M. A. Littlejohn, W. C. Holton, Scaling Trends for Device Performance and Reliability in Channel-Engineered n-MOSFET's, IEEE Transactions on Electron Devices, Vol. **ED-45**, 1998, pp. 254-260.

57. P. Packan, S. Thompson, E. Andideh, S. Yu, T. Ghani, M. Giles, J. Sandford, M. Bohr, Modeling Solid Source Boron Diffusion for Advanced Transistor Applications, International Electron Device Meeting, Technical Digest, 1998, pp. 505-508.

58. H. Hwang, D.-H. Lee, J. M. Hwang, Degradation of MOSFETs Drive Current Due to Halo Ion Implantation, International Electron Device Meeting, Technical Digest, 1996, pp. 567-569.

59. H. Hwang, D. H. Lee, J.-G. Ahn, J. S. Byun, D. Yang, Effect of Channeling of Halo Ion Implantation on Threshold Voltage Shift of Metal Oxide Field Effect Transistor, Appl. Phys. Lett., Vol. **68**, 1996, pp. 938-939.

60. M. Hargrove, S. Crowder, E. Nowak, R. Logan, L. K. Han, H. Ng, A. Ray, D. Sinitsky, P. Smeys, F. Guarin, J. Oberschmidt, E. Crabbe, D. Yee, L. Su, High-Performance Sub-0.08 µm CMOS with Dual Gate Oxide and 9.7 ps Inverter Delay, International Electron Device Meeting, Technical Digest, 1998, pp. 627-630.

61. I. Y. Yang, H. Hu, L. T. Su, V. V. Wong, M. Burkhardt, E. E. Moon, J. M. Carter, D. A. Antoniadis, H. I. Smith, High Performance Self-Aligned Sub-100 nm Metal-Oxide Semiconductor Field-Effect Transistors Using X-Ray Lithography, J. Vac. Sci. Technol. B., Vol. **12**, 1994, pp. 4051-4054.

62. S. Venkatesan, J. W. Lutze, C. Lage, W. J. Taylor, Device Drive Current Degradation Observed with Retrograde Channel Profiles, International Electron Device Meeting, Technical Digest, 1995, pp. 419-422.

63. S. E. Thompson, P. A. Packan, M. T. Bohr, Linear Versus Saturated Drive Current: Tradeoff in Super Steep Retrograde Well Engineering, Symposium on VLSI Technology, Technical Digest, 1996, pp. 154-155.

64. Y. P. Tsividis, "Operation and Modeling of the MOS Transistor", McGraw-Hill, New York, 1987.

65. J. B. Jacobs, D. Antoniadis, Channel Profile Engineering for MOSFET's with 100 nm Channel Lengths, IEEE Transactions on Electron Devices, Vol. **ED-42**, 1995, pp. 870-875.

66. A. Das, H. De, V. Misra, S. Venkatesan, S. Veeraraghavan, M. Foisy, Effects of Halo Implant on Hot Carrier Reliability of Sub-quarter Micron MOSFET's, 1998 International Reliability Physics Symposium, Proceedings, pp. 189-193.

67. D. Song, J. Lim, K. Lee, Y.-J. Park, H.-S. Min, Optimization Study of Halo Doped MOSFETs, Solid-State Electronics, Vol. **39**, 1996, pp. 923-927.

68. H. S. Wong, *et al.*, Nanoscale CMOS, Proceeding of the IEEE, Vol. **87**, 1999, pp. 537-565.

69. A. O. Adan, *et al.*, SOI Technology Status for low-power IC Applications, Proceedings of the Eighth International Symposium on Silicon-on-insulator Technology and Devices, 1997, pp. 340-351.

70. S. Asai, *et al.*, Technology Challenges for Integration Near and Below 0.1um, Proc. of IEEE, Vol. **85**, 1997, pp. 505-520.

71. T. Eimori, et al., Approaches to extra low voltage DRAM operation by SOI-DRAM, IEEE Transactions on Electron Devices, Vol. **ED-45**, 1998, pp. 1000-1009.

72. J-W. Park, *et al.*, Performance Characteristics of SOI DRAM for Low-Power Application, IEEE Journal of Solid State Circuits, Vol. **34**, 1999, pp. 1446-1453.

73. G. Shahidi *et al.*, A Room Temperature 0.1 um CMOS on SOI, IEEE Transactions on Electron Devices, Vol. **ED-41**, 1994, pp. 2405-2411.

74. F. Balestra, Double Gate Silicon-On-Insulator Transistor with Volume Inversion: A New Device with Greatly Enhanced Performance, IEEE Electron Device Letters, Vol. **EDL-8**, 1987, pp. 410-412.

75. T. Aeugle, L. Risch, W. Rosner, H. Schafer, M. Franosch, M Eller, T. Ramcke, Fabrication of Ultrashort Vertical MOS-Transistors, Proceedings of the Sixth International Symposium on Ultralarge Scale Integration Science and Technology, 1997, pp. 561-569.

76. J. Moers, D. Klaes, A. Tonnesmann, L. Vescan, S. Wickenhauser, T. Grabolla, M. Marso, P. Kordos, H. Luth, Vertical p-MOSFETs with Gate Oxide Deposition before Selective Epitaxial Growth, Solid-State Electronics, Vol. **43**, 1999, pp. 529-535.

77. H. Takato, K. Sunouchi, N. Okabe, A. Nitayama, K. Hieda, F. Horiguchi, F. Masuoka, Impact of Surrounding Gate Transistor (SGT) for Ultra-High-Density LSI's, IEEE Transactions on Electron Devices, Vol. **ED-38**, 1991, pp. 573-578.

78. L. Risch, W. Krautschneider, F. Hofmann, H. Schafer, Vertical MOS Tranistors with 70 nm Channel Length, Proceedings of the 25th European Solid State Device Research Conference, 1995, pp. 102-104.

79. S. Maeda, S. Maegawa, T. Ipposhi, H. Nishimura, H. Kuriyama, O. Tanina, Y. Inoue, T. Nishimura, N. Tsubouchi, Impact of a Vertical Φ-Shape Transistor (VΦT) Cell for 1 Gbit DRAM and Beyond, IEEE Transactions on Electron Devices, Vol. **ED-42**, 1995, 2117-2124.

80. S.-P. Jeng, R. H. Havemann, M.-C. Chang, Mater. Res. Soc. Symp. Proc., Vol. **337**, 1994, p. 25.

81. P. K. Vasudev, M. Mendicino, T. E. Seidel, Advanced Materials for Low Power Electronics, Solid-State Electronics, Vol. **39**, 1996, pp. 489-497.

82. A. S. Sedra, K. C. Smith, "Microelectronic Circuits", Oxford University Press, New York, 1998.

83. S-K. Chiang, C. L. Lassen, The Market for low-K Interlayer Dielectrics, Solid State Technology, Vol. **42**, October 1999, p. 42.

84. R. Liu, C-S. Pai, E. Martinez, Interconnect Technology Trend for Microelectronics, Solid-State Electronics, Vol. **43**, 1999, pp. 1003-1009.

85. J. R. Lloyd, J. J. Clement, Electromigration in Copper Conductors, Thin Solid Films, Vol. **262**, 1995, pp. 135-141.

86. H. B. Huntington, A. R. Grone, J. Phys. Chem. Solids, Vol. **20**, 1961, p. 76.

87. I. A. Blech, Electromigration in Thin Aluminum Films on Titanium Nitride, J. Appl. Phys., Vol. **47**, 1976, pp. 1203-1208.

88. C-K. Hu, M. B. Small, K. P. Rodbell, C. Stanis, P. Blauner, Electromigration Failure due to Interfacial Diffusion in Fine Al Alloy Lines, Appl. Phys. Lett., Vol. **62**, 1993, pp. 1023-1025.

89. C-K. Hu, B. Luther, F. B. Kaufman, J. Hummel, C. Uzoh, D. J. Pearson, Copper Interconnection Integration and reliability, Thin Solid Films, Vol. **262**, 1995, pp. 84-92.

90. C. W. Kaanta, S. G. Bombardier, W. J. Cote, W. R. Hill, G. Kerszykowski, H. S. Landis, D. J. Poindexter, C. W. Pollard, G. H. Ross, J. G. Ryan, S. Wolff, J. E. Cronin, in VMIC

1991, Proceedings of the International IEEE VLSI Multilevel Interconnection Conference, June, p. 144.

91. J. S. H. Cho, H. K. Kang, I. Asano, S. S. Wang, CVD Cu Interconnection for ULSI, International Electron Device Meeting, Technical Digest, 1992, p. 297.

92. J. Heidenreich, D. Edelstein, R. Goldblatt, W. Cote, C. Uzoh, N. Lustig, T. McDevitt, A. Stamper, A. Simon, J. Dukovic, P. Andriacacos, R. Wachnik, H. Rathore, T. Katsetos, P. McLaughlin, S. Luce, J. Slattery, Cu Dual Damascene for sub-0.25 μm CMOS, in IITC 1998, Proceedings of the International Interconnect Technology Conference, 1998, p. 151.

93. J. Hummel, G. Biery, S. Bhattacharya, S. T. Chen, Materials Issues for Copper Damascene Low K ILD Integration, Proceedings of the Low Dielectric Constant Materials and Interconnects Workshop, 1996, pp. 357-384.

94. P. C. Andricacos, C. Uzoh, J. O. Dukovic, J. Horkans, H. Deligianni, Damascene Copper Electroplating for Chip Interconnections, IBM J. Res. Develop., Vol. **42**, 1998, pp. 567-574.

95. S. P. Murarka, Low Dielectric Constant Materials for Interlayer Dielectric Applications, Solid State Technology, March 1996, p. 83.

96. L. W. Hrubesh, L. E. Keene, V. R. Latorre, Dielectric Properties of Aerogels, J. Mater. Res., Vol. **8**, 1993, pp. 1736-1741.

97. E. Korczynski, Low-k Dielectric Costs for Dual-Damascene Integration, Solid State Technology, Vol. **42**, May 1999, p. 43.

98. R. H. Haveman, K. A. Monnig, Interconnect Scaling, Performance and Reliability Challenges in the Next Millennium, Semicon West 99, Cu Interconnect Status, 1999, p. A-3.