Material and Process Limits in Silicon VLSI Technology

JAMES D. PLUMMER, FELLOW, IEEE, AND PETER B. GRIFFIN

Invited Paper

The integrated circuit (IC) industry has followed a steady path of shrinking device geometries for more than 30 years. It is widely believed that this process will continue for at least another ten years. However, there are increasingly difficult materials and technology problems to be solved over the next decade if this is to actually occur and, beyond ten years, there is great uncertainty about the ability to continue scaling metal–oxide–semiconductor field-effect transistor (MOSFET) structures. This paper describes some of the most challenging materials and process issues to be faced in the future and, where possible solutions are known, describes these potential solutions. The paper is written with the underlying assumption that the basic metal–oxide–semiconductor (MOS) transistor will remain the dominant switching device used in ICs and it further assumes that silicon will remain the dominant substrate material.

Keywords*—*Dielectric materials, MOSFETs, semiconductor device doping, semiconductor device fabrication, silicon.

I. INTRODUCTION

For more than 30 years, the integrated circuit (IC) industry has followed a steady path of constantly shrinking device geometries and increasing chip size. This strategy has been driven by the increased performance that smaller devices make possible and the increased functionality that larger chips provide. Together, these performance and functionality improvements have resulted in a history of new technology generations every two to three years, commonly referred to as "Moore's Law" [1], [2]. Each new generation has approximately doubled logic circuit density and increased performance by about 40% while quadrupling memory capacity. The increase in components per chip comes from three key factors first identified by Gordon Moore. The factor of two in component density comes from a $\sqrt{2}$ shrink in each lithography dimension. An additional factor of $\sqrt{2}$ comes from an increase in chip area and a final factor of $\sqrt{2}$ from device and circuit cleverness, providing the overall quadrupling in chip capacity. The apparent ease with which all this has happened has led to an expectation that faster and more powerful chips will continue to be introduced on the same schedule for the foreseeable future. In fact, the semiconductor industry itself has developed a "roadmap" based on exactly this idea. The National Technology Roadmap for Semiconductors (NTRS) [3] and most recently the International Technology Roadmap for Semiconductors (ITRS) [4] now extend this device scaling and increased functionality scenario to the year 2014, at which point minimum feature sizes are projected to be 35 nm and chips with $>10^{11}$ components are expected to be available. Fig. 1 summarizes the trends in feature size over time.

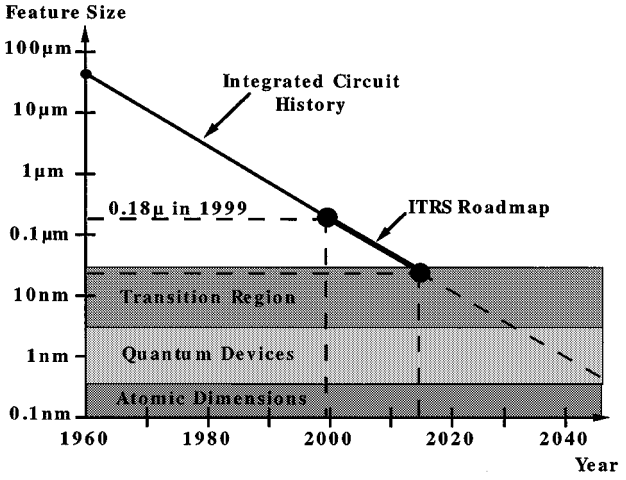


Fig. 1. Feature size versus time in silicon ICs.

Most of the history represented in Fig. 1 has been achieved with the same basic switching element (the metal–oxide–semiconductor [MOS] transistor), the same basic circuit topology (complimentary metal–oxide–semiconductor [CMOS]), and with a limited number of materials (Si, SiO_2, Al, Si_3N_4, $TiSi_2$, TiN, and W, primarily). While very substantial human and financial resources have been invested in scaling dimensions and increasing chip sizes over the past 40 years, in many respects progress in these areas

Manuscript received January 28, 2000; revised September 20, 2000.
The authors are with Stanford University, Stanford, CA 94305 USA.
Publisher Item Identifier S 0018-9219(01)02065-5.

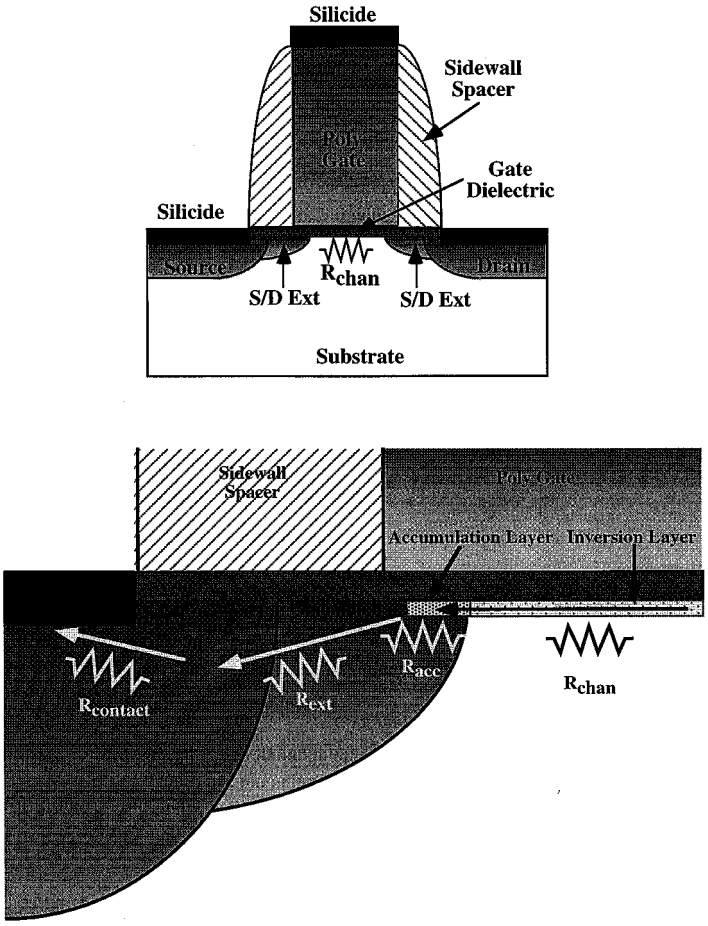


Fig. 2. Schematic cross section of a modern MOS transistor. Bottom figure is a blowup near the source/channel boundary. Arrows indicate the current flow path and the resistors illustrate the various regions that can affect the current drive capability of the device.

has been straightforward in the sense that no fundamentally new inventions have been needed. Obviously, manufacturing practices have improved. However, the device structures of 30–40 years ago and the manufacturing processes used then are quite recognizable in today's IC industry.

If the ITRS is a correct predictor of the next 15 years, superficially much will remain the same as it has for the past 30–40 years. However, there are many reasons to believe that continued device scaling will not be as straightforward in the future as it has been in the past. Practical and/or fundamental limits are being approached and substantial changes to device technologies and structures are going to be required. While "inventions" and new materials have largely not been needed for the past 30 years, they surely will be needed over the next 15 years. This period will likely be the most challenging that the IC industry has faced because it is likely that during this period we will really understand how far Moore's Law can be extended. In fact, without new materials and inventions, we will certainly see the end of Moore's Law within this period. It is likely, however, that solutions will be found to the difficult problems that lie ahead. The economic motivation is substantial and the size of today's IC industry will permit enormous resources to be applied to finding solutions.

We will describe in this paper some of the important limits and problem areas that will have to be addressed over the next 15 years. Many of these will be materials issues since it is likely that the basic MOS transistor and the basic CMOS circuit topology will remain the industry workhorses over this entire period. We will focus primarily on "frontend" issues in this paper (structures, processes, and materials associated with the switching devices in chips) since other papers in this special issue deal with backend (interconnect) issues and with higher level device, circuit, and system issues.

II. MOSFETs at the Scaling Limit—What Are the Materials Issues?

Fig. 2 schematically illustrates the basic metal–oxide–semiconductor field-effect transistor (MOSFET) device used in today's silicon chips. The basic fabrication process steps to manufacture such a device have been broadly described [5]. There are no serious competitors to replace this device in the foreseeable future. The basic structure will continue to evolve to allow continued performance improvements, but fundamental changes are unlikely in the next 15 years.

Table 1
Selected Data from the NTRS [3], ITRS [4], and Logic Technology Predictions [7]

Year of 1st DRAM Shipment	1997	1999	2002	2005	2008	2011	2014
Minimum Feature Size	250 nm	180 nm	130 nm	100 nm	70 nm	50 nm	35 nm
Isolated Gate Length	-	100 nm	70 nm	50 nm	35 nm	25 nm	18 nm
DRAM Bits/Chip	256M	1G	(3G)	8G	(24G)	64G	(192G)
DRAM Chip Size (mm^2)	280	400	460	530	630	710	860
Equivalent Physical Gate Oxide Thickness (nm)	3-5	1.9-2.5	1.5-1.9	**1.0-1.5**	**0.8-1.2**	**0.6-0.8**	**0.5-0.6**
Dielectric Constant of DRAM Capacitor		22	50	250	700	**1500**	**1500**
Max Gate Electrode Resistivity ($\mu\Omega$ cm)		60	43	33	**23**	**16**	**11**
Max Silicide/Si Contact Resistivity ρ_C ($\Omega\ cm^2$)		$30x10^{-8}$	$17x10^{-8}$	$\mathbf{10x10^{-8}}$	$\mathbf{5x10^{-8}}$	$\mathbf{2.5x10^{-8}}$	$\mathbf{1.5x10^{-8}}$
S/D Extension Sheet Resistance (Ω/sq)		350-800	250-700	**200-625**	**150-525**	**120-450**	**100-400**
S/D Extension x_j (nm)	50-100	42-70	25-43	**20-33**	**16-26**	**11-19**	**8-13**
S/D Extension Lateral Abruptness (nm/decade)		14	8.5	6.5	**4.5**	**3.2**	**2.2**
Minimum Supply Voltage (volts)	1.8-2.5	1.5-1.8	1.2-1.5	0.9-1.2	0.6-0.9	0.5-0.6	0.5

Various scenarios have been proposed for scaling the MOSFET device, the simplest of which is due to Dennard *et al.* [6]. The "ideal" scaling they proposed maintains constant electric fields in the device by shrinking all voltages, currents, and physical dimensions by the same factor λ (typically $\sqrt{2}$ in each generation) and increasing all doping concentrations by the same factor. Actual scaling scenarios followed by the semiconductor industry have not shrunk voltages and currents as rapidly as physical dimensions with the result that electric fields have increased over time. The motivation for doing this is simply the higher device performance achievable (more current drive) when electric fields are increased. It is likely that these general trends will continue and, in fact, they are the basis for the NTRS and ITRS projections. Selected data from these roadmaps [3], [4] and logic technology predictions [7] are shown in Table 1.

There are a number of issues associated with continued MOSFET scaling that represent challenges for the future and, ultimately, fundamental limits. The bold entries in Table 1 correspond to requirements for which there are currently no known solutions or at least no solutions that generally are believed will work in manufacturing. These entries provide the main topics for this paper.

The first issue is the gate dielectric thickness. By purely geometric arguments, the gate insulator in a MOSFET needs to be thin compared to the device channel length in order for the gate to exert dominant control over the channel potential. This avoids "short channel effects," which are largely the result of the drain electric field penetrating throughout the channel and influencing the channel potential at the source side of the device. Practical MOSFET structures generally require the gate dielectric thickness to be a few percent of the channel length. The entries in Table 1 for gate dielec-

tric thickness are unrealizable starting in 2005 if SiO_2 is the dielectric material. Oxides thinner than about 1.0–1.5 nm conduct direct-tunneling currents too large for most IC applications at the supply voltages listed in the table. A new higher dielectric constant material system will have to be employed starting with the 100-nm generation. A "higher k" material system will allow a physically thicker dielectric layer to have an "equivalent SiO_2 thickness" corresponding to the entries in Table 1. Higher k dielectric materials are also needed for dynamic random access memory (DRAM) storage capacitors. However, the requirements are quite different in this application because only a charge storage function is required. The two-dimensional (2-D) effects in a logic device that arise because of the difference in permittivity between the silicon channel and the gate insulator make the dielectric requirements very different in logic devices [8]. Generally, the highest dielectric constant possible is needed for DRAMs capacitors as shown in Table 1, whereas materials with dielectric constants up to 30 or so are needed for active transistor gate insulators.

The gate electrode itself also presents some significant challenges. Polysilicon has been used for more than 25 years as the gate electrode material. However, decreasing its resistivity, as shown in Table 1, implies increasing the doping levels in the polysilicon, which minimizes the resistivity of the gate electrode and helps avoid polysilicon depletion effects. But this approach is limited by dopant solubility limits and by dopant outdiffusion from the poly through the thin gate dielectric and into the silicon. This later problem is particularly acute with P^+ gates because boron diffuses rapidly through SiO_2. The likely solution is again new materials—metal gate electrodes. But there are no known materials solutions that are known to work in manufacturing.

The next two rows in Table 1 relate to parasitic resistances in the MOSFET. Ideally, the current drive in a MOSFET is limited by the intrinsic channel resistance (R_{chan} in Fig. 2). In practice, all the other resistances in Fig. 2 play a significant role and degrade the intrinsic device capability. Normal design procedures require these other resistances to total less than 10% of the channel resistance. Such requirements determine the entries in Table 1 for contact resistivity and for source/drain extension sheet resistance. Both rows show bold entries beginning in 2005. Contacts are almost always made with either $TiSi_2$ or $CoSi_2$ contacting heavily doped silicon today. There is no known manufacturable means to reduce the contact resistance (R_{contact} in Fig. 2) in these systems to the values specified in Table 1. The entries for source/drain extension sheet resistance (R_{ext} in Fig. 2) are also in bold starting in 2005. Here the issue is that the junction depths of these extensions (x_J in Fig. 2) must continue to decrease to minimize short channel effects. Thus, doping levels in these regions must increase in order to keep resistances low. But doping concentrations are limited by dopant solubility and, hence, there are lower bounds on achievable sheet resistances for a given x_J. It is possible to incorporate metastable doping concentrations well above normal solubilities by laser melting. But there are currently no known methods to maintain these metastable concentrations through the normal heat cycles required for device fabrication.

The resistance of the extension region is also partially determined by the formation of a surface accumulation region (R_{acc} in Fig. 2) that forms under the gate in the tail region of the source/drain profile. This resistance is strongly affected by the abruptness of the extension profile because the steeper the profile is, the shorter this accumulation region will be. Thus, the entries in Table 1 for abruptness indicate progressively sharper profiles as the technology progresses.

The entries in Table 1 related to the doping profile for the source/drain extension regions (x_J and profile abruptness) are also determined by short channel effects in the device. Such effects are minimized by shallow and very abrupt junction profiles. Thus, the depth decreases and the slope becomes steeper in these profiles in the out years of the roadmap. The sheet resistance and depth entries are bold starting in 2005 and the slope entries become bold in 2008. The difficulty in fabricating shallow and very steep profiles arises largely because ion implantation is the assumed doping technique and because of this, an anneal must be performed to repair implant damage. During this anneal, transient enhanced diffusion (TED) dominates dopant diffusion, enhancing dopant diffusivities by orders of magnitude. This makes it very difficult to keep dopant profiles shallow and steep.

III. Gate Insulators

Of all of the issues outlined in the previous section, the one that requires the nearest term attention is the scaling of the gate dielectric. The interface between silicon and its native oxide SiO_2 is atomically abrupt and electrically perfect to first order. This is somewhat surprising because most semiconductor/insulator interfaces do not have this combination of properties. Even in the Si–SiO_2 case, there is a large volume expansion (2.2×) that takes place when SiO_2 is thermally grown on silicon, resulting in a highly stressed interface with the oxide under compressive stress. Yet the Si–SiO_2 system is able to accommodate these stresses and produce a virtually perfect electrical interface with trap and fixed charge densities corresponding to less than one surface defect in 10^5 surface silicon atoms. MOS transistors require this level of interface perfection and, thus, any replacement for the Si–SiO_2 system will need to at least approach these standards.

The entries in Table 1 for gate dielectric thickness are unrealizable starting in 2005 if pure SiO_2 is the dielectric material. Quantum mechanical tunneling of carriers through a barrier increases exponentially with decreasing insulator thickness. Oxides thinner than 1.0–1.5 nm operating at 1 V conduct direct-tunneling currents too large to accommodate standby power requirements in most IC applications. Thus, a new higher dielectric constant material system will have to be employed starting with the 100-nm generation. In the limit of thin oxides, the wave functions of the gate and the

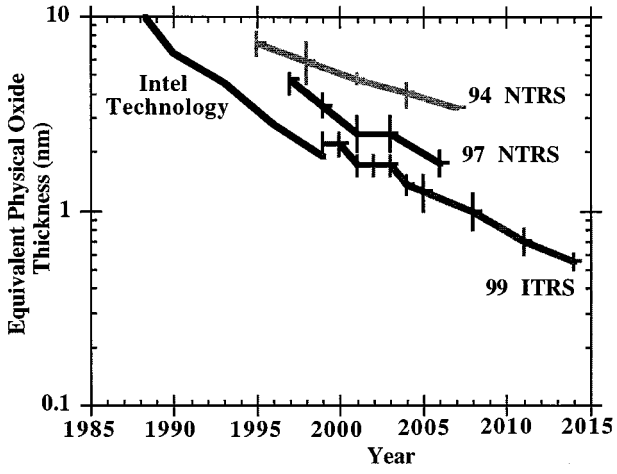


Fig. 3. ITRS predictions and Intel logic technology data for gate oxide thickness versus time.

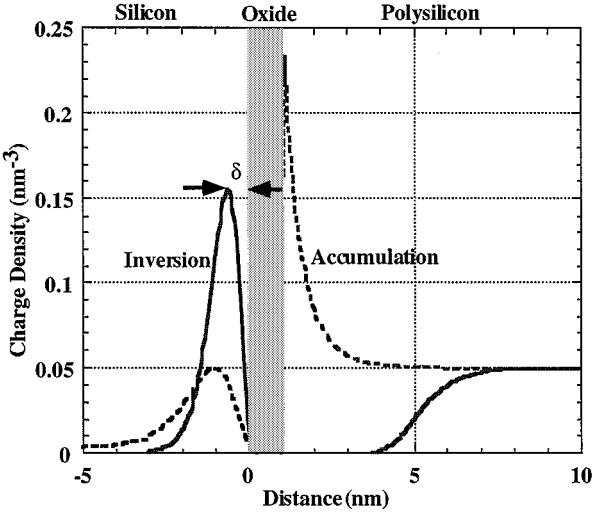


Fig. 4. Quantum calculation of the inversion (solid lines) and accumulation (dashed lines) charge distributions in silicon for a 1.0-nm oxide layer at bias voltages of 2 V using NEMO [11]. Polysilicon depletion effect is seen in inversion and the peak electron and hole charge in the silicon is below the oxide interface because of quantum confinement effects.

silicon substrate begin to overlap, causing scattering and reduced mobility. This is predicted to occur below 1.0 nm when the oxide is approximately five atomic layers thick [9], [10].

The data from the ITRS along with high-performance logic technology data are plotted in Fig. 3 and show just how quickly the oxide thickness limits are being reached. The data in Fig. 3 is given in terms of the equivalent physical oxide thickness, which represents the physical thickness of the gate dielectric corrected by the ratio of the gate dielectric constant relative to silicon dioxide. Thus, a silicon nitride gate dielectric ($\varepsilon_r = 7.8$) can be almost twice as thick as a silicon dioxide gate dielectric ($\varepsilon_r = 3.9$) and still provide the same electrical coupling of the gate to the channel.

Not all of the applied gate voltage is efficiently coupled to the channel because of polysilicon depletion effects in the gate electrode and quantum confinement effects in the silicon substrate. The diagram in Fig. 4 indicates that the equivalent electrical gate capacitance is composed of a combination of the depletion capacitance in the polysilicon, the physical dielectric capacitance, and the contribution from quantum effects in the silicon that shift the peak of the carrier distribution away from the interface. The combination of poly depletion and the quantum correction add approximately 0.8 nm to the equivalent physical oxide thickness to give the equivalent electrical thickness of the dielectric in the gate capacitor.

The polysilicon depletion effect occurs when the device is biased toward inversion and some of the applied voltage begins to deplete the highly doped polysilicon near its interface with the gate dielectric. Increasing the polysilicon active doping, especially near this interface, helps minimize the depletion effect allowing more of the applied gate voltage to influence the channel. The ITRS calls for active poly doping of 2.2×10^{20} in the 180-nm node, rising to 1.2×10^{21} at the 35-nm node. Typical maximum activation levels in polysilicon material near the dielectric interface are less than the activation levels in the crystalline source/drain, perhaps because of grain boundaries in the polycrystalline material. It is unlikely that the high activation levels called for in the ITRS can be obtained in polysilicon gates, though *in situ* doped material and polycrystalline silicon germanium material show improved activation levels. Thus, to better couple the applied gate voltage to the channel region, metal gates that are immune to the depletion effect become an attractive option.

The quantum effects in the silicon are fundamental and occur because of carrier quantum confinement/exclusion in a potential field. A quantum calculation of the charge distribution using the computer code NEMO [11] is shown in Fig. 4. The carriers drop to a low value at the barrier interfaces and the peak of the carrier distribution moves deeper into the silicon. The contribution to the effective dielectric capacitance depends on $\varepsilon_{\rm si}/\delta$, where $\varepsilon_{\rm si}$ is the silicon dielectric constant (11.7) and δ is the effective distance of the carriers below the interface, as indicated in Fig. 4.

Many high dielectric constant materials react with silicon and, therefore, need a silicon dioxide buffer layer or interface layer between the silicon and the high k dielectric. The total capacitance is then given by

$$\frac{1}{C_{\rm eff}} = \frac{1}{C_{\rm ox}} + \frac{1}{C_{{\rm hi}-k}} \tag{1}$$

and is dominated by the capacitance of the low-k material (the oxide buffer layer). In terms of the thickness and dielectric constant of the material, the effective physical oxide thickness becomes

$$T_{\rm ox}^{\rm eff,\,phys} = T_{\rm ox} + \frac{\varepsilon_{\rm ox}}{\varepsilon_{{\rm hi}-k}} T_{{\rm hi}-k}. \tag{2}$$

The effective electrical thickness of the dielectric, which is what determines the capacitive coupling between the gate and the channel, is then

$$T_{\rm ox}^{\rm eff,\,elec} = T_{\rm ox}^{\rm eff,\,phys} + T_{\rm qm} + T_{\rm pd} \tag{3}$$

where $T_{\rm qm}$ is the quantum correction for charge in the channel ($\sim$0.3 nm) and $T_{\rm pd}$ is the correction for poly depletion ($\sim$0.5 nm).

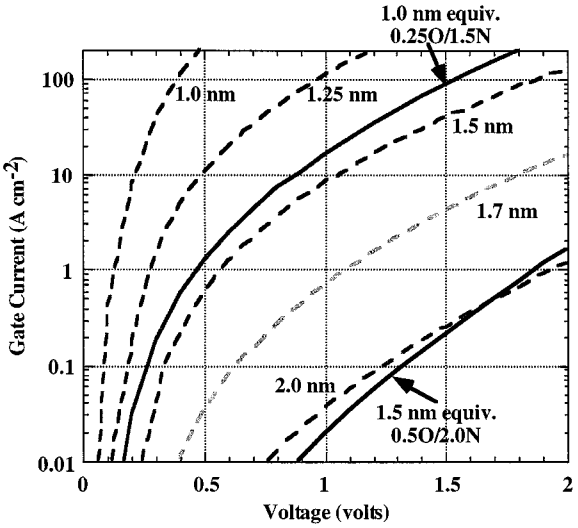


Fig. 5. Tunneling currents for oxide thickness from 1.0–2.0 nm (dashed lines) versus voltage for an NMOS device in the ON state, calculated using NEMO [11]. Solid lines represent calculations of the leakage current for a 1.0 nm (comprised of a 0.25-nm oxide layer and 1.5-nm nitride layer) and 1.5 nm (comprised of a 0.5-nm oxide layer and a 2.0-nm nitride layer) equivalent physical oxide thickness, showing dramatic improvements in the leakage current.

Fig. 5 shows calculated direct-tunneling currents for various physical oxide thicknesses at typical bias voltages for a negative-channel metal–oxide semiconductor (NMOS) device in the ON state. Direct-tunneling current in a MOS device depends on the combination of tunneling probability and the number of tunneling carriers. Because of the higher oxide tunneling barrier for holes than for electrons and the heavier effective mass of holes, the hole tunneling current is approximately an order of magnitude less than the electron tunneling current at any bias condition.

When the gate is positively biased and an NMOS transistor is in the ON state, tunneling occurs between the inversion channel electrons and the polysilicon gate. If we assume that the drain current in a 100-nm device is approximately 1 mA/μm, we can constrain the gate current to be perhaps 1% of that without seriously impacting the gain of the device. This constrains the area leakage to be

$$1\% \times 1\ \text{mA}\ \mu\text{m}^{-1}/0.1\ \mu\text{m} = 10^4\ \text{Acm}^{-2}. \qquad (4)$$

Thus, even extremely high levels of gate leakage current will not measurably affect the drive current performance of small devices. At these extreme gate-current levels, the reliability of the thin oxides is of more concern [12].

If we assume that a tolerable gate leakage is 100 Acm^{-2} [7] and assume that the gate area is 1% of a 1-cm^2 chip with a power-supply voltage of 1 V, the power dissipation due to the gate current is 1 W. This does not change significantly with temperature, while the device off current increases by a factor of a 100 between room temperature and 100°C and can approach 10% of the active power. Thus, in the highest performance applications where the active power dissipation will be more than 100 W, the high-gate tunneling currents may not be a real issue until the power dissipation due to the gate leakage exceeds the power dissipation due to the off current. In the highest performance applications, active cooling will be used, which has been shown to be able to dissipate 1000 W cm^{-2} on silicon chips [13].

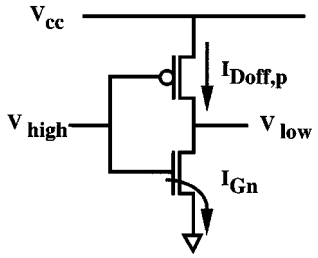


Fig. 6. Schematic of standby power dissipation current paths in a CMOS circuit configuration with the NMOS device biased in the ON state.

Another way to examine the gate-current problem is to consider that it is constrained to contribute no more than the normal off current of a device in the standby mode in a CMOS configuration. As shown in Fig. 6, in a CMOS gate in standby mode with the NMOS gate biased high, an inversion layer exists even though the drain node is discharged and the source is connected to ground. Tunneling of electrons from the inversion channel to the positively biased gate contributes a gate current that should not exceed the off current in the positive-channel metal–oxide–semiconductor (PMOS) device. In a 100-nm device with an off current of 100 nA μm^{-1}, this constrains the gate-current density to be 100 A cm^{-2} [14], [15]

$$\begin{aligned} I_{\text{gate}} &= (10^2\ \text{Acm}^{-2})(0.1 \times 10^{-4}\ \text{cm}) \\ &= 10^{-3}\ \text{Acm}^{-1} \\ &= 100\ \text{nA}\ \mu\text{m}^{-1} \approx I_{\text{off}}. \end{aligned} \qquad (5)$$

Even if the standby power requirements are relaxed by an order of magnitude, this limits the gate oxide thickness to a value greater than 1 nm.

Direct-tunneling currents that flow continuously in thin oxides pose some reliability concerns. Defects such as traps and interface states gradually build up in the oxide to a point where the oxide suddenly and destructively breaks down [16]. Because of the exponentially increasing tunneling current with decreasing oxide thickness, the time to breakdown decreases unless the gate voltage is sufficiently reduced. Avoiding reliability concerns has the same solutions as decreasing the standby power consumption. A material with a higher dielectric constant, which does not introduce more traps, can increase the time to breakdown by decreasing the gate current.

The tunneling current depends on the effective mass in the dielectric, the barrier height, and the barrier thickness, with the tunneling probability for the case of a rectangular barrier being given by [17]

$$T = \exp\left(-2\sqrt{\frac{2\,m^* q\phi_B}{\hbar^2}} \cdot t_{\text{ox}}\right). \qquad (6)$$

As the bandgap of the insulator increases, the dielectric constant of binary oxides that might be deposited on silicon tends

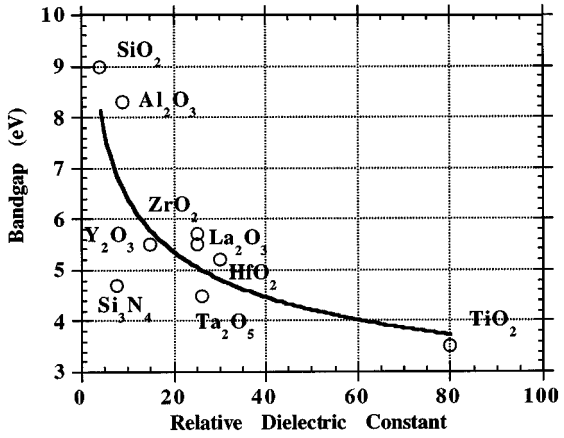


Fig. 7. Bandgap versus dielectric constant for simple binary oxides [20].

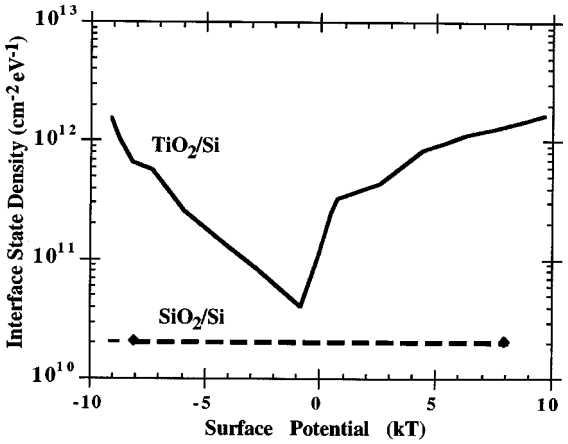


Fig. 8. Interface state density for a potential replacement gate dielectric [20].

to decrease, as shown in Fig. 7. In general, a large bandgap is desirable since the barrier height generally scales with the bandgap. Because of the square-root dependence on the barrier height and the linear dependence on thickness, there is usually an advantage in choosing a higher dielectric constant material when the aim is to reduce the tunneling current.

The precise band lineup at the silicon dielectric interface is particularly important for insulators with a smaller bandgap than SiO_2. To avoid thermal emission over a Schottky barrier, a barrier height of more than 1 eV is needed for both positive and negative carriers. Silicon dioxide has a large enough barrier and amply meets this requirement with a barrier of 3.1 eV for electrons and 4.7 eV for holes. In general, the band lineup can be quite asymmetrical, which poses potential problems for smaller bandgap materials [18]. For example, Ta_2O_5 has a bandgap of 4.4 eV, but only has an electron barrier of 0.36 eV to silicon, but a 2.9 eV barrier for holes [18]. Thus, Schottky emission of electrons into the conduction band is predicted to be a significant cause of leakage for Ta_2O_5 films on silicon. There should also be few traps in any alternative gate dielectric or leakage currents can increase due to trap-assisted tunneling, which lowers the effective tunneling barrier.

Another major barrier any new dielectric material will have to overcome is to achieve almost the same low-defect density that occurs at the interface between silicon and silicon dioxide. This is because any improvement in the drive current caused by better capacitive coupling to the gate can easily be lost if the carrier mobility degrades. There is an approximate relationship between mobility and interface state density that was empirically found for thermal oxide field-effect transistors (FETs) [19]

$$\frac{\mu}{\mu_o} = \frac{1}{1 + k\overline{D}_{IT}} \tag{7}$$

where $\overline{D}_{IT}(\mathrm{cm}^{-2}\ \mathrm{eV}^{-1})$ is the concentration of charged states at the bias condition. This is the integral of the interface state density between the surface potential corresponding to the onset of strong inversion and the band edge. For example, the interface state density for TiO_2 in contact with silicon has been reported to be $3 \times 10^{10}\ \mathrm{cm}^{-2}$ at midgap, rising with a "V" shape toward the band edges rather than with the "U" shape typically seen in thermal oxide capacitors [20], as shown in Fig. 8. This was shown to be enough to reduce the mobility by a factor of two. Thus, a low value of midgap interface state density is no guarantee that mobility will not be degraded under device operating conditions.

Because of the large number of possible gate dielectric materials, systematic approaches to suggesting or eliminating candidates are useful. One of the most comprehensive studies of the thermodynamic stability of potential binary oxides was carried out by Hubbard and Schlom [21]. Its main usefulness is in predicting when an interfacial buffer layer is likely to be needed for high k dielectrics. Attempts to predict the dielectric constant of the high-k materials based on simple theory have been largely unsuccessful. Another general approach to estimating the interfacial properties between a dielectric and silicon is based on considerations of bonding constraints and coordination number [22]. This study predicted comparable interface state densities for silicon nitride and for titanium, tantalum, and aluminum oxides, all higher than for silicon dioxide.

Because silicon nitride has approximately twice the dielectric constant of oxide, it is an attractive material for increasing the gate dielectric constant [23]. Gate oxides today are heavily nitrided in an NO or N_2O ambient to reduce boron penetration and to increase reliability. However, the nitrogen levels at the interface are closely controlled and are in the range of one atomic percent. Higher levels of nitrogen incorporation can lead to degraded device characteristics. These oxynitride gate dielectrics do not have a measurably different dielectric constant than a pure SiO_2 gate oxide. In order to raise the dielectric constant, it is necessary to use a nitride rich or pure silicon nitride layer [24]. It is known that silicon nitride directly in contact with silicon has a large interface state density, which degrades device performance. For this reason, a thin interfacial oxide rich layer is needed with a nitride dielectric. We can estimate the improvement that a gate stack consisting of an oxide buffer layer and a nitride dielectric provides as follows. Fig. 9 shows the relevant parameters for the calculation [25].

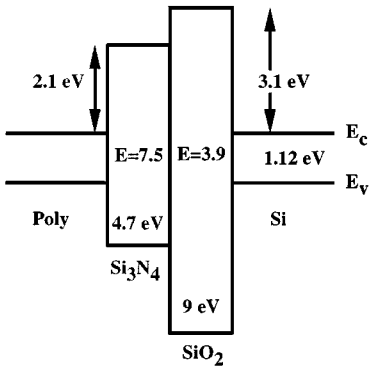


Fig. 9. Nitride and oxide parameters important for determining barrier tunneling currents.

A simple estimate of the improvement in the tunneling current in a stacked dielectric can be obtained as follows [15]. An equivalent physical oxide thickness of 1.5 nm can be obtained from a 0.5-nm buffer oxide interface layer and a 2.0-nm-thick nitride layer for a physical thickness of 2.5 nm

$$t_{\rm ox}^{\rm eff,\,phys} = 0.5 + \left(\frac{3.9}{7.8}\right) 2.0 = 1.5 \text{ nm}. \tag{8}$$

Because the tunneling probability for the composite stack based on (6) is proportional to

$$\sqrt{\phi_b^{\rm ox}} t_{\rm ox} + \sqrt{\phi_b^{\rm nit}} t_{\rm nit} \tag{9}$$

the composite layer with a 0.5-nm SiO_2 buffer layer and a 2.0-nm Si_3N_4 layer will have a tunneling probability similar to that of a 2.1-nm physical SiO_2 layer. Thus, the gain from a nitride layer amounts to 0.6-nm equivalent physical oxide thickness for this stack. This assumes that the effective mass in both dielectrics is the same. Full quantum calculations of the improvements for stacks with 1.0- and 1.5-nm effective physical oxide thickness are shown in Fig. 5. The advantage of the higher k layer with the SiO_2 buffer layer diminishes further as the thickness of the stack decreases because the buffer layer must remain at a monolayer or more of oxide. However, every angstrom gained represents a significant improvement in the leakage current, so any advances in improving the dielectric constant of the gate stack will be important.

This simple example shows the difficulty of incorporating a buffer layer in a gate stack with an equivalent physical oxide thickness of less than 1.0 nm. It also shows the importance of the barrier height for electrons in the band lineup between the gate electrode and the dielectric material. Because the bandgap tends to decrease with higher dielectric constant for simple oxides, any asymmetric band lineup can cause leakage problems due both to direct tunneling and thermionic emission over the barrier. Asymmetric band lineups have been predicted for several high-k dielectric materials in contact with silicon, including tantalum oxide, strontium titanate, and barium strontium titanate [18].

The periodic table in Fig. 10 indicates many of the popular choices for high-k gate dielectrics that have been experimentally investigated, along with the binary oxides predicted to be stable in contact with silicon at 1000°K from thermodynamic calculations or experimental review [21].

Titanium and tantalum oxides have been proposed for high-k gate dielectrics [26]–[28]. Thermodynamically, neither oxide is stable in contact with silicon. Titanium tends to form the metal silicide while tantalum prefers the metal phase

$$3\text{Si} + \text{TiO}_2 \rightarrow \text{TiSi}_2 + \text{SiO}_2 \quad \Delta\text{G} = -23 \text{ kcal/mol} \tag{10}$$

$$\tfrac{5}{2}\text{Si} + \text{Ta}_2\text{O}_5 \rightarrow 2\text{Ta} + \tfrac{5}{2}\text{SiO}_2 \quad \Delta\text{G} = -52 \text{ kcal/mol}. \tag{11}$$

A complex ternary interfacial oxide composed of Si–Ta_x–O_y, which converted to SiO_2 after annealing, was found at the tantalum oxide/silicon interface [29], [30]. Because these oxides are thermodynamically unstable in contact with silicon, an SiO_2 barrier layer must be used. High-leakage currents are reported with polysilicon gates [20], so if polysilicon gates are used, barrier layers must be formed at both interfaces to produce good results [31]. When the object is to form a layer equivalent to a 1.0-nm physical oxide, this leaves little process margin.

Oxides of hafnium and zirconium are medium-k materials that are thermodynamically predicted to be stable in contact with silicon. Hafnium and zirconium can be thought of as the transition metal counterparts of the column 4 tetrahedral solids and their oxides have a structure like SiO_2. Since they are elements from column D4 of the periodic table with four electrons in d states, there are just enough electrons to replace the silicon. In practice, sputtered material forms an interfacial layer upon annealing which appears to be a silicate layer [32], [33]. Upon higher temperature annealing, the silicate layer converts to a more stoichiometric oxide layer and the films tend to crystallize. The silicate interfacial layer appears to be important in reducing the interface state density [32], though it does add a layer with a lower dielectric constant. Because of the reactions seen, it is likely that metal gates would need to be used in combination with these metal oxides, adding enormously to the process integration issues. A replacement gate process is also a possibility [34].

Lanthanum and yttrium oxides are column III dieclectrics that have received attention as medium-k materials. Aluminum oxide is another medium-k material with a bandgap near SiO_2 that has been investigated [35], [36]. Many other binary oxides have been proposed, but it seems unlikely that simply depositing an oxide on silicon will result in a structure as stable and perfect as a thermal SiO_2–Si interface.

Based on the desirable properties of the medium-k hafnium and zirconium oxides, the hafnium and zirconium silicates (Zr, Hf)Si_xO_y have been investigated as gate dielectrics [37]. Alloying the metal oxides with SiO_2 or equivalently doping SiO_2 with elements with high polarizability is expected to improve the interface quality over that of the metal oxides, making it more SiO_2-like in terms of band alignment, dangling bonds, and trap densities.

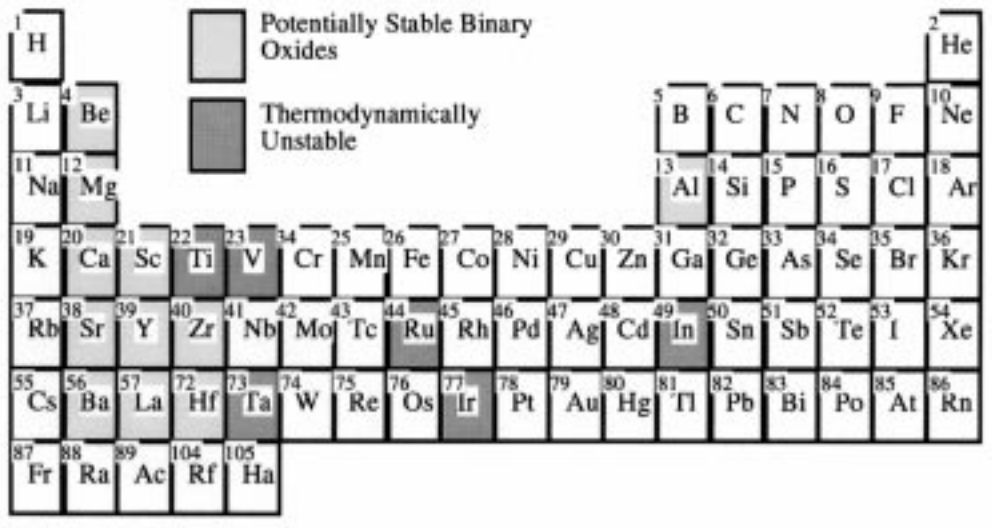


Fig. 10. Periodic table of elements with possible dielectric choices indicated [21].

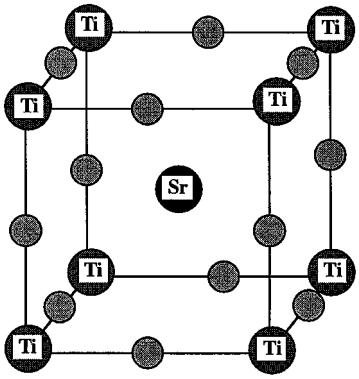


Fig. 11. Structure of perovskite material.

The crystalline structure for $ZrSiO_4$ is a body-centered tetragonal composed of SiO_4 tetrahedra interspersed with Zr atoms. Hafnium silicate is expected to have the same structure. The use of a silicate composed of SiO_2 and ZrO_2 or HfO_2 structural units should act more like the SiO_2–Si interface. The dielectric constants of the silicates are lower than the metal oxides, but are intermediate between the low value of SiO_2 and the higher values for the metal oxides.

A very important class of materials are the mixed metal oxide materials called perovskites, which have the formula ABC_3, where the C is oxygen. They constitute one of the largest structural families known in solid-state chemistry because many substitutions of the metal cations are possible, including replacement of a cation by a combination of cations of different valency such that the net charge remains identical. The perovskites are of interest because they may allow epitaxial crystalline oxides in perfect registry with the silicon substrate [38]. Strontium titanate is a simple perovskite that has been widely investigated for capacitor applications in memory cells and its structure is shown in Fig. 11. The perovskite structure of $SrTiO_3$ is composed of a simple cubic lattice of titanium ions with oxygen ions at the center of every cube edge. The Sr ions reside at the cube centers and their major function is providing electrons to the system. Strontium has two electrons outside its krypton-like core while titanium has four electrons outside its argon-like core; these six valence electrons compensate the six negative charges of the three oxygen atoms that compose the $SrTiO_3$ structure. The Sr conduction band states are very high in energy and are of no importance in determining the electronic structure of the material [39]. The electronic structure and interface gap states are determined by the Ti–O bonds [18]. Based on the primary role of Sr in contributing electrons to the system, it is expected that Ba or Ca substitutions would behave very similarly.

Predictions for systems controlled by Ti–O or Ta–O bonds are that Schottky barrier pinning is sizable and leads to a low-conduction band offset [18]. The thermodynamic stability of Ti–O- or Ta–O-based materials has already been shown to be poor in contact with silicon. Indeed, the barium–strontium–titanate/silicon interface is too reactive to be used unless a low-temperature processing route is available; in applications for DRAM capacitor structures, it is sandwiched between two metals.

In spite of these problems, a thermodynamically stable binary oxide such as SrO or BaO can act as a buffer layer between the silicon surface and the high-K perovskite or an interfacial silicide can act as a growth template [38]. The attraction of the layered perovskite materials is that they might provide an epitaxial route to growing a high-k layer on silicon. To do so, it is likely that many of the same tricks used to grow compound heterostructures will be used. Because of the various transition metals that can be substituted in the perovskite structure, a lattice match with silicon is possible.

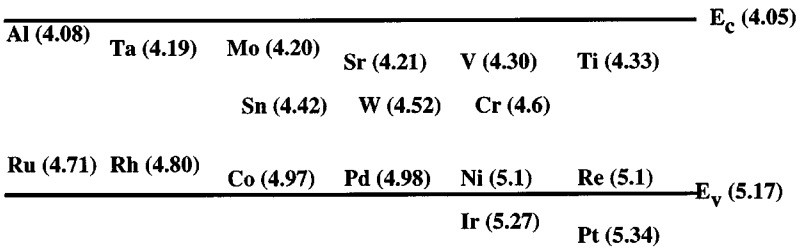


Fig. 12. Workfunctions of possible metal gates.

Complex perovskites such as (Ba, Sr)La_n(Sc, Al)$_n O_{3n+1}$ provide a large window of possible opportunities when combined with an underlying buffer layer.

IV. Gate Electrodes

Many of the new dielectric materials are unstable in direct contact with silicon and by extension are unstable in the presence of the polysilicon gate material also. Thus, it is likely that the entire gate stack will have to be replaced, with metal gates replacing polysilicon. Polysilicon has the advantage that it can be doped p-type or n-type, shifting the workfunction so that it is suitable for NMOS and PMOS devices. A compromise midgap workfunction metal gate shifts the threshold voltage higher by half the bandgap (about 0.5 V) compared to a highly doped polysilicon gate electrode. If the channel doping is lowered to provide a suitable lower threshold voltage, then the channel doping is too low to control short-channel effects. For this reason, two different gate metals are required with workfunctions near E_C and E_V. The incorporation of two different metal gates in a CMOS flow enormously complicates the fabrication process.

There are several candidate metals with workfunctions near E_C: Al, Ta, Mo, Zr, Hf, V, Ti, and several candidate metals with workfunctions near E_V: Co, Pd, Ni, Re, Ir, Pt. The diagram in Fig. 12 shows the workfunction of the common metals in the silicon bandgap. There are also several conducting metal oxides as one moves toward the right side of the periodic table in the transition metal oxides: In_2O_3, SnO_2, OsO_2, RuO_2, IrO_2, ZnO, MoO_2, ReO_2, and conducting nitrides such as TiN. The problem is not a lack of choices, but rather, which choice can be cost effectively integrated into the process flow. If one abandons a conventional process flow and moves toward a gate last process flow, then even the constraints of thermal stability in the gate stack are lessened. With the expertise available in chemical mechanical polishing (CMP), alternative process flows may provide a viable path toward integrating new gate stacks.

The combination of metal gates and an oxide/nitride gate dielectric allows an extension by one or two technology generations beyond what is possible with polysilicon and SiO_2. It is interesting to note that any medium k dielectric replacement for oxide in high performance applications might already have to demonstrate reliability at current levels of 1 Acm^{-2}, once thought to be the limit for silicon dioxide. Given the trends in Moore's Law, a likely scenario is that device and circuit cleverness will be used to control the power dissipation problems caused by high-gate leakage. Multiple thicknesses of gate oxides will be used on a chip to manage power and reliability constraints. Oxides will continue to be used in the highest performance circuits because of the accumulated knowledge about their reliability behavior. A gradual introduction of heavily nitrided oxides, oxide/nitride stacks, or medium-k dielectrics will be used to lower leakage currents in most of the chip until knowledge of their reliability behavior accumulates. Given the enormous investment in silicon technology, it is likely that these challenges will be met even if they seem very difficult at present. There appear to be no fundamental materials limits on scaling the gate stack at least for the next decade.

V. Shallow Junctions

The issues associated with shallow junctions were briefly introduced in connection with Fig. 2. There are two issues which dominate the constraints placed on the source/drain junction extension regions, parasitic resistance, and short channel effects. Short channel effects are a result of the drain electric field extending through the channel region and therefore modulating the channel potential near the source and are discussed in more detail in other papers in this special issue [40].

The parasitic resistance issue is quite simple. The MOS transistor operates through the gate potential modulating the charge in the channel (inversion) region. The magnitude of the inversion charge and the length of the channel determine the resistance associated with this region in the device. Simple first-order MOS device physics gives the channel resistance as

$$R_{\text{chan}} = \left[\frac{W}{L}\,\mu\,\frac{\varepsilon_{\text{ox}}}{t_{\text{ox}}}\,(V_{\text{gs}} - V_{\text{th}})\right]^{-1} \tag{12}$$

where

- W and L gate width and channel length, respectively;
- V_{gs} applied gate voltage (usually the supply voltage);
- V_{th} threshold voltage.

As device geometries are scaled down, ideal scaling [6] suggests that W, L, t_{ox}, V_{gs}, and V_{th} all decrease at the same rate. In this scenario then, R_{chan} would remain constant as the technology is scaled. Therefore, if this were done, the various parasitic resistances in Fig. 2 would simply need to remain relatively constant in value from generation to generation.

Because higher performance has been a specific objective of scaling the technology, ideal scaling has not been fol-

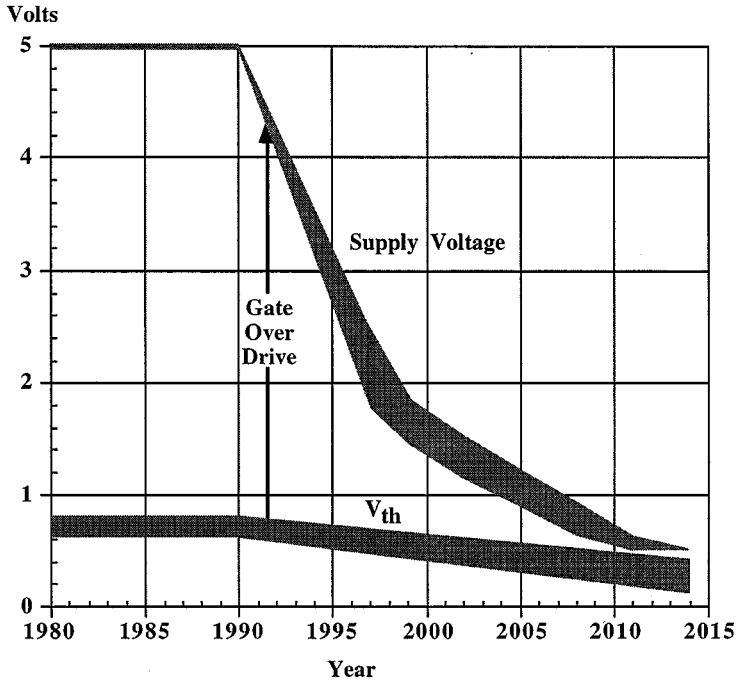


Fig. 13. Supply voltage and $V_{\rm th}$ versus time from the NTRS and ITRS.

lowed in the past and will not likely be followed in the future. Higher performance is achieved by higher current drive capability and, therefore, $R_{\rm chan}$ has been decreased as the technology has been scaled. This has been achieved by aggressively shrinking $t_{\rm ox}$ (see Fig. 3) and by shrinking $V_{\rm gs}$ and $V_{\rm th}$ more slowly than ideal scaling would suggest. The latter point is illustrated in Fig. 13. Note that these voltages were not scaled at all until the 1990s with the result that $R_{\rm chan}$ decreased approximately proportionally with feature size until supply voltage scaling began. Thus, over time, the parasitic resistances have also been required to decrease in order to keep $R_{\rm chan}$ as the dominant resistive component in the device.

Although these topics are covered in more detail in other papers in this special issue [40], it is worthwhile making some comments about Fig. 13 because of the impact this figure has on materials issues. The fact that the supply voltage has not been scaled as rapidly as ideal scaling might suggest has resulted in increased electric fields in device structures. Thus, materials, especially the gate dielectric, have been pushed closer to materials limits. The vertical electric fields in gate dielectrics in today's MOS transistors are typically greater than 5 MV cm^{-1} and may reach twice this value toward the end of the ITRS. These values are approaching the physical limits of SiO_2 and most other dielectric materials.

Towards the end of the roadmap, the supply voltage is projected to be about 0.5 V and is limited by the maximum sustainable fields in the very thin gate dielectrics at that time and by power consumption issues in the very large chips expected to be in manufacturing at that time. It should be noted in Fig. 13 that there is considerable uncertainty about the device threshold voltages expected to be used at that time. The threshold voltage typically needs to be a fraction of the supply voltage in order to provide sufficient gate drive ($V_{\rm gs}$–$V_{\rm th}$) for high-speed performance. Yet the threshold voltage also needs to be well above 0 V in order to minimize the off state leakage current in the transistors. The rate of change of device current with gate voltage is fundamentally limited at room temperature to a 10× change in current for each 60-mV change in gate voltage because of the mechanism of current flow (thermionic emission over a potential barrier). Thus, the threshold voltage typically needs to be several times 60 mV above 0 V in order to achieve reasonable off state currents. In today's devices, this is achieved with $V_{\rm th}$ values of about 0.5 V. With a supply voltage of 1.5–3 V, this provides reasonable gate drive in the ON state. However, with a supply voltage of 0.5 V in 2014, there is no room left for designing device $V_{\rm th}$ to simultaneously meet OFF-state and ON-state current requirements. This issue is perhaps the most serious "device" issue facing future generations of technology. Innovations such as electronically controlled $V_{\rm th}$ voltages will likely be needed and are discussed elsewhere in this special issue [40].

When current leaves the channel inversion layer (Fig. 2), it first flows through an accumulation region before entering the source/drain extension region. The lateral profile of this doped region is a key issue. The doping in the extension drops from a peak concentration of 10^{19} cm^{-3} today (10^{20} cm^{-3} at the end of the roadmap) to the background channel doping 10^{17} cm^{-3} today ($>10^{18}$ cm^{-3} at the end of the roadmap) over a distance determined by the profile slope. The abruptness of this slope is specified in Table 1. At least over the latter part of the profile the extension doping is lower than the carrier concentration induced in the surface accumulation layer by the gate. Thus, the current flows through the accumulation layer until it reaches a region in the extension where the doping exceeds the accumulation layer carrier den-

sity. Typically, the carrier density in the accumulation region is on the order of 10^{18}–10^{19} cm^{-3} [41], so the lateral extent of the accumulation region can be significant depending on the extension region lateral doping profile.

When the current leaves the accumulation region, it does so gradually, resulting in a spreading resistance R_{ext} in series with the accumulation layer resistance R_{acc}. Because of the complexity of the doping profiles in modern devices, these resistances are best calculated today using numerical simulation, but there are simple analytic expressions available that do provide useful estimates. Ng and Lynch [41] derived such expressions and showed that both R_{acc} and R_{ext} strongly depend on the steepness of the doping profile in the extension region. Physically, this is because the length of the accumulation region is determined by how steep the profile is and, hence, R_{acc} decreases as the slope becomes steeper. R_{ext} is also reduced by a steeper profile because the current quickly moves into more highly doped (lower resistance) regions when the profile is steep. The entries in Table 1 relating to the abruptness of the extension profile are a result of these considerations.

In Fig. 2, the gate is shown overlapping the source/drain extension by a significant distance. In the context of the above discussion, it is clear why this is required. The gate forms and controls the accumulation region at the surface of the extension and, hence, controls R_{acc}. If the gate overlap is too short then the accumulation region cannot extend as far as it otherwise might, with the result that R_{ext} may be significantly increased. There are a number of reports in the literature of decreases in device performance when this gate overlap is reduced too far [42]. The required gate overlap can of course be reduced by increasing the steepness of the extension profile. In the limit of a perfectly abrupt junction, the required overlap would be reduced to just the lateral extent of the junction depletion region in the extension.

The best one can do in terms of profiles is a perfectly flat constant-concentration abrupt (box-shaped) profile whose sheet resistance is given by

$$\rho_S = \frac{1}{q\mu C_S x_J} \tag{13}$$

where C_S is the surface concentration and x_J is the junction depth. The box-shaped profile encloses the maximum dose for a given x_J and, hence, represents the minimum sheet resistance. x_J values are specified in Table 1 and are determined primarily by short channel effects. C_S is limited by electrical solubility of the dopant atoms in the silicon crystal and the mobility μ is then the value appropriate for silicon doped at a concentration C_S.

The maximum concentration of a dopant that can be dissolved in silicon under equilibrium conditions without forming a separate phase is termed the solid solubility [43]. Many of the elements used as dopants exhibit a retrograde solid solubility, where the maximum concentration that can be dissolved occurs below the melting point as shown in Fig. 14. It is the electrically active concentration that is most important to device designers. Though the solid solubility

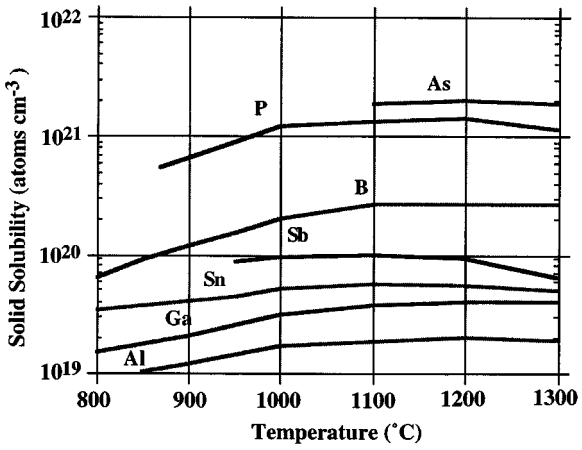


Fig. 14. Solid solubility curves for various dopants in silicon. Values are the equilibrium solubilities at each temperature and may not be achieved in device doped regions (after [43]).

is the thermodynamic maximum concentration that can be accommodated in a solid without a separate phase forming, kinetic effects may limit the electrically active dopant concentration that can be achieved under typical processing conditions. By this we mean that if the wafer temperature is changed, some time is required for the dopant solubility to reach the value characteristic of the new temperature. In addition, the electrical solubility limit may be considerably lower than the maximum solid solubility shown in Fig. 14 because of neutral cluster formation with point defects in the silicon lattice. Typically, dopants above the electrical solubility limit form an inactive complex that is electrically neutral and does not contribute free carriers to the doped region.

Solid solubility data like Fig. 14 suggests that arsenic might be active up to concentrations of 2×10^{21} cm^{-3}, but in practice it is difficult to actually achieve electrically active arsenic concentrations above 2×10^{20} cm^{-3} [42], [44]. The origin of this discrepancy is of enormous practical interest. It is true that techniques such as laser melting of the silicon can introduce arsenic into silicon in metastable electrically active concentrations near the solubility limit. However, there is an enormous driving force that tends to deactivate the arsenic during any subsequent thermal cycling. Upon annealing, some of the arsenic, while not strictly forming a separate precipitate phase, forms an electrically inactive structure. One such proposed structure, which is consistent with the experimental evidence, is that of several arsenic atoms surrounding a vacancy. The arsenic atoms remain on substitutional sites, but adjoin a vacancy that leaves the arsenic three-fold coordinated with the silicon lattice while retaining two electrons in a dangling bond for a full shell of eight electrons. Thus, the As atoms are not electrically active in this form and do not contribute free electrons to the crystal.

If we use the ITRS x_J specifications and assume the limiting active doping concentration is 2×10^{20} cm^{-3} and the limiting carrier mobility is 52 cm^2 V^{-1} s^{-1} (which is a reasonable value for both n- and p-type silicon [45]), then we

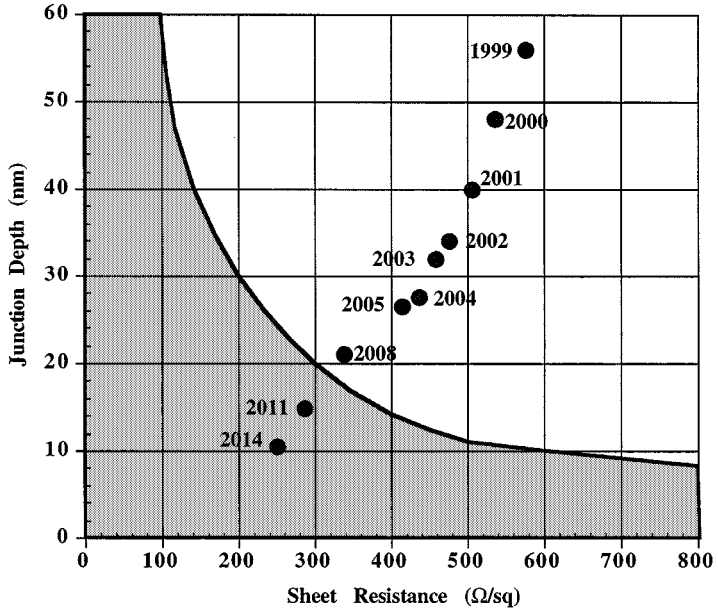


Fig. 15. Source/drain extension junction depth and sheet resistance requirements from the ITRS [4]. Given the x_J values shown, sheet resistances in the dotted area are excluded if a dopant solubility of 2×10^{20} cm^{-3}, a mobility of 52 cm^2 V^{-1} s^{-1}, and an ideal box profile are assumed.

can construct the plot shown in Fig. 15. One can immediately see that the ITRS requirements beyond 2008 (the 70-nm node) cannot be met even by an ideal box-shaped profile if the doping concentration is limited to the electrical solubility. R_{ext} is the resistance most affected by the extension region ρ_S. Thus, this suggests that the ITRS requirements for R_{ext} cannot be met for the last few technology generations if the device structure remains as shown in Fig. 2 and C_S is limited to equilibrium electrical solubility values.

The situation is actually worse than this because the techniques used today to form doped regions in MOS devices (ion implantation followed by a rapid thermal anneal to activate the dopants) do not even come close to an ideal box-shaped profile. Practical profiles achievable with today's manufacturing techniques are typically a factor of two to three worse than the ideal profile in terms of achievable ρ_S because such profiles are graded, which implies that the ITRS requirements will not be achievable perhaps as soon as the 130- or 100-nm generations. This is why bold entries appear in Table 1 for ρ_S at the 100-nm generation. We will return to this point later.

There may be opportunities in this area for new conceptual approaches. One possibility is the use of metastably doped silicon, that is, the incorporation of doping concentrations in excess of the normal solubility limits. It is well established that such doping levels can be achieved by laser annealing for example [44]. In these processes, the silicon is locally melted and dopants can be "frozen in" at electrically active concentrations above 10^{21} cm^{-3} during the very rapid cooling that occurs after the laser pulsing. Such doping concentrations are metastable, however, and any subsequent heat cycles provide a huge driving force for precipitation and deactivation of the dopants.

VI. Junction Contacts

The final component of device resistance shown in Fig. 2 is the contact resistance. Contacts in today's device structures are normally made by self-aligned silicides contacting heavily doped silicon. This process provides an ohmic contact completely covering the area of the source/drain diffusions and, therefore, minimizes the contact resistance.

Current flows in a distributed manner from the source/drain extension to the contact. The exact flow lines depend on the doping profile in the silicon and on the geometry. The effective contact resistance depends on this flow pattern or, in other words, on the effective area of the contact. Current crowding on the leading edge of the contact can be a significant effect. In this structure, the contact resistance is given by [46]

$$R_{\text{contact}} = \frac{\sqrt{\rho_c R_{sd}}}{W} \coth(L_c/L_t) \approx \frac{\rho_c}{W L_c} \qquad (14)$$

where

- ρ_c specific contact resistivity of the silicide/semiconductor contact (Ω cm^2);
- R_{sd} sheet resistance of the source/drain diffusion (Ω/square);
- W contact width;
- L_c contact length.

$L_t = \sqrt{\rho_c/R_{sd}}$ is called the transfer length and is the average distance that carriers travel in the diffusion before entering the contact. For typical values of ρ_c and R_{sd}, L_t is greater than the physical contact length L_c, which results in the approximation in (14). In this case, the current flows into the entire length of the contact and current crowding

effects are minimal. Thus, the contact resistance varies inversely with the contact area if ρ_c is constant.

The silicide formation process itself often consumes silicon since the metal component (Ti, for example) is usually deposited and then reacted to form the silicide. This has several important consequences. First, some of the volume of the heavily doped source/drain regions is lost or consumed by the silicide formation. The portion of the source/drain region which is "lost" is the top portion, which is normally the most heavily doped and, therefore, the most conductive—exacerbating the resistance problem. This increases the sheet resistance of the remaining diffusion in which current can flow to the contact and, therefore, increases the effective contact resistance. Finally, the resistivity of an ohmic connection between a metal or silicide and heavily doped silicon depends strongly on the doping level adjacent to the metal. Silicon consumption and dopant segregation behavior during the silicide formation process can strongly affect this doping level.

For a tunneling contact, the specific contact resistivity depends on the semiconductor-metal barrier height and the silicon doping [46]

$$\rho_c = \rho_{c0} \exp\left(\frac{2\sqrt{\varepsilon_s m^*}}{\hbar} \frac{\phi_B}{\sqrt{C_S}}\right) \tag{15}$$

where

- ρ_{c0} contact resistivity for an infinitely high active surface doping concentration;
- C_S actual active surface doping concentration;
- ϕ_B barrier height between the silicon and metal or silicide.

ρ_{c0} is process dependent and can change with cleaning or contact etching procedures.

Required values for ρ_c in Table 1 are derived from a desire to limit the total parasitic resistance in the device structure to no more than 10% of the channel resistance. If the area of the silicide/semiconductor contact is taken to be a square defined by the minimum feature size, then the implied contact resistance from the numbers in Table 1 is approximately 1000 Ω, independent of technology generation. With reducing feature sizes, this constant resistance is achieved by requiring the contact resistivity ρ_c to scale over time directly with the contact area.

Based on (15) two key parameters in reducing ρ_c are the silicon doping C_S and the barrier height ϕ_B. C_S is limited today by the electrical solubility of dopants in the silicon, as discussed earlier. If we take the barrier height as one half of the silicon bandgap (0.55 eV) and assume that the maximum electrically active dopant concentration in the silicon is 2×10^{20} cm^{-3}, then ρ_c is limited to about 1×10^{-7} Ω cm^2. Thus, the required ρ_c values beyond 2005 are not achievable by the contacting schemes currently employed in CMOS technology. New approaches will be needed. The obvious areas to focus on are the barrier height ϕ_B and the dopant solubility C_S.

Barrier height engineering in the metal–silicon system is often thought to be impractical because the barrier height is found to be largely independent of the metal workfunction due to Fermi-level pinning. At the same time, barrier heights to silicides range from approximately 0.4 to 0.9 eV, though little is known about the workfunction of silicides. At a fundamental level, the formation of a Schottky contact to a semiconductor is not fully understood and barrier height engineering might be possible by incorporation of thin interfacial layers between the metal or silicide and the semiconductor to modify the barrier properties. This expectation is based on the role of interface charges, electronegativity, bond strength, and dipole moments in determining the barrier height.

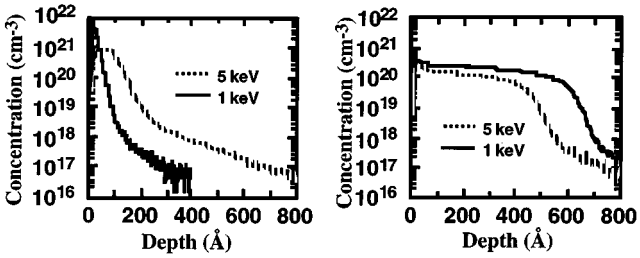


Fig. 16. Measured (SIMS) profiles of a 5-keV and a 1-keV 1×10^{15} cm^{-2} arsenic implant. (a) As-implanted. (b) After a 1050°C 10-s RTA anneal. 5-keV profile shows 30% dose loss [49].

The other parameter C_S may be more amenable to new conceptual approaches. One possibility is the use of metastably doped silicon, as described earlier. The near-noble-metal silicides tend to form at relatively low temperatures (200–600°C) while the refractory metal silicides form at higher temperatures (>600°C). Because of the importance of maintaining any metastable active doping concentration in the silicon during further thermal processing, this may put more emphasis on ultralow temperature silicide formation processes in the future. Some silicides such as $CoSi_2$ and $NiSi_2$ are interesting because they have the calcium fluorite crystal structure with a lattice constant that is very close to silicon. These silicides are candidates for low temperature epitaxial growth, a process which has not yet been widely investigated. Limitations that often occur because of the different behavior of n- and p-type dopants during the conventional silicide formation process might be avoided by an admittedly more complex epitaxial growth process. An additional possibility is to form the silicide before the dopants are implanted and then implant directly into the silicide, followed by outdiffusion from the silicide into the silicon to form the junctions. This strategy eliminates the doped silicon consumption that normally occurs during silicide formation and may be useful in reducing ρ_c [47].

These and other possibilities will have to be explored in the near future if the requirements of the ITRS for ρ_c are going to be met. It is not clear at this point how "fundamental" a limiting ρ_c value of 10^{-7} Ω cm^2 is. If this is a practical or fundamental limit, new innovations will be required in device structures to continue the performance improvements of the past and these innovations will be required in the very near future.

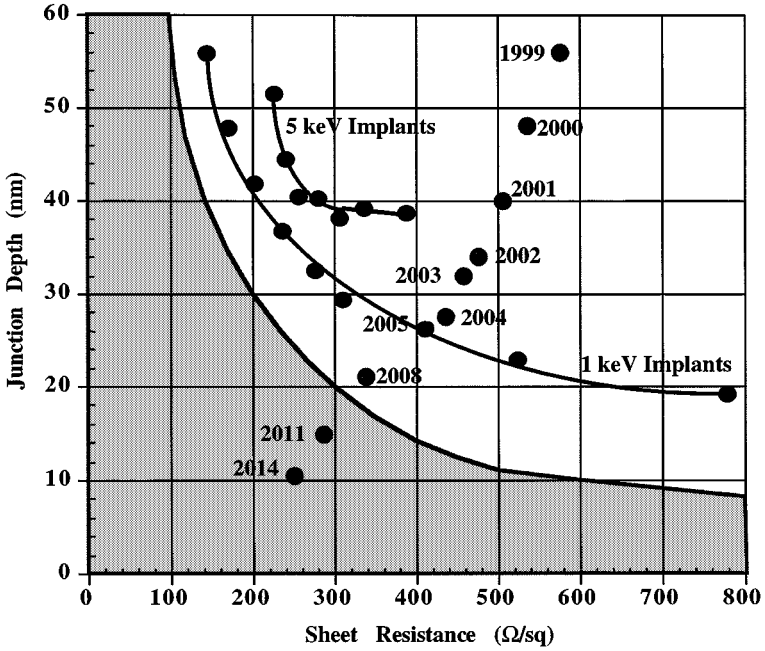


Fig. 17. Experimental results are superimposed on Fig. 5. In both the 1- and 5-keV implants, 1×10^{15} cm^{-2} doses and RTA anneals were used [49].

VII. Junction Formation—Ion Implantation and Annealing

The dominant technology used today for doping silicon is ion implantation. This process provides precise control of the placement and quantity of doping atoms. In this process, doping ions (e.g., As^+) are accelerated to energies of 1–1000 keV and implanted into the silicon using photoresist or hard masks like SiO_2 to block the implant where it is not desired. The range of these ions in silicon varies inversely with energy and so the need for shallower junctions in scaled devices generally implies lower implant energies. The stopping process of these ions in the silicon involves both nuclear collisions and electronic drag forces, both statistical processes, with the result that a doping profile is produced by the implant. Examples are shown in Fig. 16 for low-energy As implants.

The implantation process produces considerable damage in the crystalline silicon substrate as a result of the nuclear collisions involved in the stopping process. At the high-dose levels typical of implants used to form source and drain regions or extension regions, the silicon is generally turned amorphous. At an atomic level, the damage consists of isolated point defects (interstitials and vacancies) small complexes or clusters of these point defects and more extended defect structures.

Dopants in silicon diffuse by interaction with point defects. In the crystalline state, silicon contains small equilibrium concentrations of isolated point defects. These concentrations depend exponentially on temperature and are typically close to zero at room temperature but 10^8–10^{14} cm^{-3} at process temperatures (700–1100°C). Dopant atoms diffuse by pairing with either vacancies or interstitials and hopping from lattice site to lattice site in the crystal [48]. In implanted silicon, the damage created by the implant creates large excess point defect concentrations which lead to anomalous dopant diffusion while the damage is being repaired. This repair process essentially occurs by providing thermal energy to the lattice so that silicon atoms can find their way back to lattice sites. This "anneal" can be done at moderate temperatures for long times (700°C for tens of minutes to hours) or at high temperatures for short times (1000°C for 10 s). Either way, significant anomalous dopant diffusion takes place during the anneal. This diffusion is known as TED and is a major issue in making shallow junctions. Fig. 16 also illustrates the broadening of the implanted profiles during a typical high-temperature rapid thermal anneal (RTA). TED is generally minimized by using a short high-temperature anneal rather than a long low-temperature anneal essentially because the required anneal time decreases faster with temperature than diffusivities increase with temperature.

In the example in Fig. 16, it is also interesting to note that the lower energy implant actually diffuses past the higher energy implant during the anneal step. This seems completely counterintuitive but can be understood because the higher energy implant loses 30% of its dose due to segregation to the Si/SiO_2 interface during the anneal. Dopant diffusivities are very concentration dependent, so the lower concentrations in the 5-keV implant diffuse more slowly than the higher concentrations in the 1-keV implant. The mechanisms underlying all these effects are only beginning to be fully understood and it is not clear today why the dose loss occurs in the 5-keV case and not in the 1-keV case [49]. There are many other possible implant process conditions which need to be explored to provide answers to these questions.

Based on experiments of the type shown in Fig. 16, various combinations of implant energies and anneal cycles have been carried out to assess whether the requirements specified in the ITRS are achievable or not. One example is shown

in Fig. 17 in which experimental results for low-energy implants and RTA anneals are superimposed on the ITRS requirements in Fig. 15. Using 1-keV implants and optimum anneals, the 2005 (100 nm) node of the ITRS is barely achievable and the last three nodes of the roadmap are outside the range of possibility at least for the conditions of these experiments. Even the ideal profile cannot achieve the requirements in the last one or two technology nodes of the ITRS as was pointed out in the earlier discussion associated with Fig. 15. Note that the 1-keV implant and optimum anneal data come within about a factor of two of the ideal box profile limit, suggesting that there is not much additional room for improvement if these shallow junctions are limited by the doping solubility.

At very high concentrations, dopant diffusivities increase dramatically, making it more difficult to keep junctions shallow under reasonable anneal conditions. This increase in dopant diffusivity is partly due to concentration dependent diffusivities. The mechanism behind these effects is believed to be the idea that point defects (vacancies and interstitials) can exist in both neutral and charged states in the silicon lattice. The charged defect concentrations are governed by the energy levels of these defects in the silicon bandgap and by the Fermi level (E_F) [48], [50]. In highly doped material, the Fermi level moves away from its intrinsic level in the middle of the bandgap and can cross over the levels corresponding to the charged point defects when the doping is high enough. This results in exponential increases in the concentrations of these defects and, hence, exponential increases in dopant diffusivities as E_F moves. The result is dopant diffusivities that depend on n/n_i or p/n_i if negatively charged or positively charged point defects are involved respectively. The dependencies go as $(n/n_i)^2$ or $(p/n_i)^2$ if doubly charged defects are dominant. While these effects do increase dopant diffusivities and, hence, make obtaining shallow junctions more difficult, they also have the positive effect of making dopant profiles more box-like because the high-concentration regions diffuse more rapidly and therefore produce a flatter profile near the surface.

Other anomalous diffusion effects also occur at high concentrations, many of which are poorly understood today. In boron doped regions, diffusivities have been found to be even larger than concentration dependent diffusivities would predict. This has been termed boron-enhanced diffusion or BED and is believed to be due to the formation and diffusion of additional boron defect pairs that form at high concentrations [51].

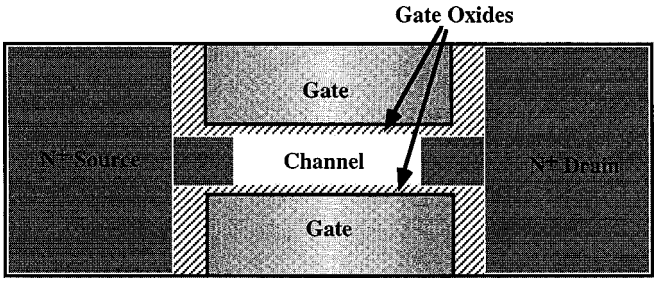
TED effects can also be greatly affected when the junctions are shallow. Experimentally, it is found that anomalous diffusion due to implant damage is actually less of an issue when the junctions are very shallow [52]. This is likely because with a surface or material interface nearby, the large excess concentrations of point defects created by the implant process can recombine rather than diffuse into the bulk where they can affect dopant diffusion. Surfaces and interfaces are thought to be very effective sites for excess point defect recombination, although this probably depends on the kind of interface (Si-SiO_2 or some other interface) and on the condition of the surface (whether it has been damaged by a previous implant or not). Considerably more work is needed in these areas to fully understand all the issues.

Perhaps the most promising opportunity to radically alter the limits, shown in Fig. 17, is to find some way to exceed equilibrium solid solubility limits for dopant concentrations. As described earlier, metastable doping concentrations in excess of 10^{21} cm^{-3} can easily be achieved by laser melting. Solidification of the silicon is so rapid following this process that dopant concentrations corresponding to the solubility near the melt temperature are frozen into the crystal. As also described earlier, however, such dopant concentrations easily deactivate under subsequent thermal cycling. Thermal cycles currently used (and believed to be necessary for future device processing) are likely sufficient to deactivate enough dopants that the final active concentrations will be no better than today's limit of about 2×10^{20} cm^{-3}. However, this conclusion is not certain and needs to be more carefully investigated. Deactivation involves the clustering of dopant atoms, typically with point defects. If methods can be found to suppress the formation of these point defects during any necessary anneals, perhaps the metastable doping concentrations can be maintained. Alternatively, or perhaps in combination with these methods, device processing perhaps can be redesigned to limit thermal cycles so that deactivation is less of a problem. Finally, an alternative device structure may be found that circumvents the issue completely by relaxing the constraints on parasitic resistances shown in Fig. 2.

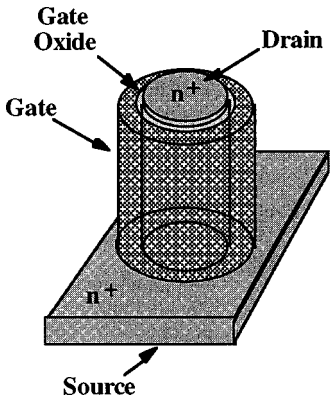
VIII. MOSFETs at the Scaling Limit

The two most obvious materials issues in scaled MOSFETs at the end of the ITRS are the gate insulator and the need for ultrashallow junctions. The first of these does not seem very amenable to solution by alternate device structures because the MOSFET depends fundamentally on a gate electrode separated from the channel region by a high-quality insulator. The second issue, shallow junctions, is less fundamental and can be attacked somewhat by device structure innovation. We will explore both of these issues further in this section.

Consider first the gate insulator issue. The ITRS specifications for insulator thickness are driven primarily by two issues: short channel effects and device performance. Short channel effects are primarily a geometric issue. The gate needs to have tighter control of the channel potential than the drain does in order to avoid modulation of the drain current by the drain voltage. Conventionally, this has implied that the gate insulator thickness needs to be a few percent of the channel length. The ITRS assumes that this ratio will be maintained in the future and this is what leads to insulator thicknesses of <1 nm (equivalent SiO_2 thickness) when the channel length is 35 nm. The most obvious solution to this problem is to change to a higher dielectric constant insulator as described earlier. But this solution has very difficult materials issues, especially if it is pushed beyond the oxynitride class of materials. In fact, it is very likely that an Si/SiO_2 interface at least a monolayer thick will always be required for



(a)



(b)

Fig. 18. (a) Schematic representations of double-gate and (b) surround-gate MOS transistor structures.

high-quality MOSFET devices, which places a lower limit on insulator thickness of about 0.3 nm even if tunneling currents were not a problem.

Some relief from this problem can be obtained by device structure innovations. A number of groups are exploring double-gate [53] and surround-gate MOS [54] structures today. These structures, illustrated conceptually in Fig. 18, essentially configure the MOS gate on two sides (top and bottom) of the channel in the case of the double-gate structure or completely around the channel like a sheath around a wire in the case of the surround-gate structure. Either structure affords better gate control of the channel potential and, thus, for a given channel length, they relax the gate insulator thickness specification. Since current flows between drain and source on both sides of the channel (double gate) or completely around the perimeter of the silicon "wire" (surround gate), these structures also improve the performance of the basic MOSFET device. These device structures should allow scaling by an additional one or two generations if the ITRS scenario is followed, but they do not, of course, solve the fundamental problem of a limit on how thin the gate insulator can be made. In some respects, they actually complicate this problem because the gate insulator now has to be grown or deposited on a more complex geometry and this may provide additional practical limits on how thin the gate insulator can be.

Another approach that has been proposed [55] suggests simply accepting 1.5 nm as the limit on insulator scaling, using SiO_2 as the insulator since it works at these dimensions and setting the supply voltage at 1 V and $V_{\rm th}$ at 0.2 V since these values provide reasonable leakage currents and

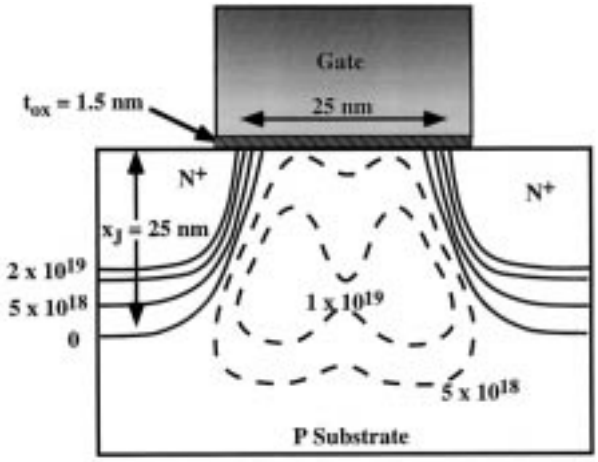


Fig. 19. MOSFET design at 25-nm dimensions using a different scaling scenario than the ITRS (after [55].

gate drive. The question then is whether such a device design can be modified to make it work properly at small dimensions. By modifying the lateral channel doping profile as illustrated in Fig. 19, reasonable device performance was predicted at effective channel lengths of 25 nm, toward the end of the ITRS. These complex lateral doping profiles require careful ion implantation and suppression of TED effects so that the profiles stay where they are implanted. This may not be feasible in such very small structures. Alternatively, vertical MOSFET structures like the surround-gate device in Fig. 18 could use epitaxial growth to produce complex channel doping profiles. This approach might be more feasible for implementing structures like that shown in Fig. 19.

The series resistance issues described earlier in connection with the device structure in Fig. 2 present serious challenges as shown for example in Fig. 17. This problem is perhaps amenable to solution again through device structure innovation. The ITRS sets the junction depth x_J at 10 nm, and it is this combination of x_J and ρ_S that is unachievable in Fig. 17. However, the structure illustrated in Fig. 19 has a junction depth of 25 nm and in the simulations described in [55], x_J values as high as 50 nm were used, with reasonable short channel characteristics. The combination of a 25- or 50-nm x_J and the sheet resistance values in the ITRS in 2014 is achievable by standard ion implantation and RTA techniques since these values fall well above the dotted forbidden region in Fig. 17. Thus, the ITRS specifications may in fact be achievable even at the end of the roadmap through continued device innovation. Other device structures including raised source/drain configurations have also been proposed to help address the series resistance problem [56].

We have not specifically addressed other proposed device options such as silicon-on-insulator (SOI) devices or the incorporation of heterostructures using materials such as SiGe. These innovations do not change the fundamental materials issues addressed in this paper with respect to conventional CMOS technology. They may provide additional design options or more performance in some applications, but these issues are outside the scope of this paper. Other papers in this special issue address some of these topics.

IX. Conclusion

The silicon semiconductor industry has charted a course for itself over the next 15 years, which basically continues the density and performance improvements of the past 40 years. There do not appear to be many, if any, "fundamental" materials issues that will prevent this from actually occurring. Nevertheless, these future technology generations outlined in the ITRS will not come as easily as past generations have. Silicon technology has reached the point at which significant materials innovations will be required to circumvent very significant challenges. The most apparent of these are the need for new gate insulator materials and achieving low-sheet resistance ultrashallow junctions. It may well also be the case that the basic planar MOS transistor structure will have to be modified in significant ways, perhaps through adoption of vertical or double-gate structures, to help extend device scaling toward the end of this 15-year period. Nevertheless, because of the enormous economic incentive to continue density and performance improvements, the most likely scenario for the future is exactly the scenario of the past, device scaling along with innovation and creativity to continue the silicon CMOS microelectronics revolution.

References

[1] G. E. Moore, "Cramming more circuits on chips," *Electron.*, vol. 19, pp. 114–117, Apr. 1965.

[2] ——, "Progress in digital integrated electronics," in *Tech. Dig. 1975 Int. Electron Devices Meeting*, Dec. 1975, pp. 11–13.

[3] Semiconductor Industry Association, *National Technology Roadmap for Semiconductors*. San Jose, CA: SIA, 1997.

[4] Semiconductor Industry Association, *International Technology Roadmap for Semiconductors*. San Jose, CA: SIA, 1999.

[5] J. D. Plummer, M. D. Deal, and P. B. Griffin, *Silicon VLSI Technology—Fundamentals Practice and Modeling*. Englewood Cliffs, NJ: Prentice-Hall, 2000.

[6] R. H. Dennard, F. H. Gaensslen, H.-N. Yu, V. L. Ridout, E. Bassous, and A. R. LeBlanc, "Design of ion-implanted MOSFETs with very small physical dimensions," *IEEE J. Solid-State Circuits*, vol. 9, p. 256, Oct. 1974.

[7] T. Ghani, K. Mistry, P. Packan, S. Thompson, M. Stettler, S. Tyagi, and M. Bohr, "Scaling challenges and device design requirements for high performance sub-50nm gate length planar CMOS transistors," in *Tech. Dig. 2000 Symp. VLSI Technol.*, June 2000, pp. 174–175.

[8] D. J. Frank, Y. Taur, and H.-S. P. Wong, "Generalized scale length for two-dimensional effects in MOSFETs," *IEEE Electron Device Lett.*, vol. 19, p. 385, Oct. 1998.

[9] D. A. Muller, T. Sorsch, S. Moccio, F. H. Baumann, K. Evans-Lutterodt, and G. Timp, "The electronic structure at the atomic scale of ultrathin gate oxides," *Nature*, vol. 399, p. 758, June 1999.

[10] M. S. Krishnan, L. Chang, T. J. King, J. Bokor, and C. Hu, "MOSFETs with 9 to 13A thick gate oxides," in *Proc. Int. Electron Devices Meeting*, Dec. 1999, p. 241.

[11] D. K. Blanks *et al.*, "NEMO: General release of a new comprehensive quantum device simulator," in *Proc. IEEE 24th Int. Symp. Compound Semiconductors*, Sept. 1997, pp. 639–642.

[12] N. Yang, W. K. Henson, and J. J. Wortman, "Analysis of tunneling currents and reliability of NMOSFETs with sub-2nm gate oxides," in *Proc. Int. Electron Devices Meeting*, Dec. 1999, p. 453.

[13] D. B. Tuckerman and R. F. W. Pease, "High-performance heat sinking for VLSI," *IEEE Electron Device Lett.*, vol. 2, p. 126, May 1981.

[14] S.-H. Lo, D. A. Buchanan, Y. Taur, and W. Wang, "Quantum-mechanical modeling of electron tunneling current from the inversion layer of ultra-thin oxide nMOSFETs," *IEEE Electron Device Lett.*, vol. 18, p. 209, May 1997.

[15] S. E. Thompson, private communication.

[16] S. Takagi, M. Takayanagi, and A. Toriumi, "Experimental examination of physical model for direct tunneling current in unstressed/stressed ultrathin gate oxides," in *Proc. Int. Electron Devices Meeting*, Dec. 1999, p. 461.

[17] K. K. Ng, *Complete Guide to Semiconductor Devices*. New York: McGraw-Hill, 1995.

[18] J. Robertson and C. W. Chen, "Schottky barrier heights of tantalum oxide, barium strontium titanate, lead titanate and strontium bismuth tantalate," *Appl. Phys. Lett.*, vol. 74, p. 1168, Feb. 1999.

[19] S. C. Sun and J. D. Plummer, "Electron mobility in inversion and accumulation layers on thermally oxidized silicon surfaces," *IEEE Trans. Electron Devices*, vol. 27, p. 1497, Aug. 1980.

[20] S. A. Campbell, D. C. Gilmer, X. C. Wang, M. T. Hsieh, H. S. Kim, W. L. Gladferter, and J. Yan, "MOSFET transistors fabricated with high permitivity TiO_2 dielectrics," *IEEE Trans. Electron Devices*, vol. 44, p. 104, Jan. 1997.

[21] K. J. Hubbard and D. G. Schlom, "Thermodynamic stability of binary oxides in contact with silicon," *J. Mater. Res.*, vol. 11, no. 6, p. 2757, Nov. 1996.

[22] G. Lucovsky, Y. Wu, H. Nimmi, V. Misra, and J. C. Phillips, "Bonding constraints and defect formation at interfaces between crystalline silicon and advanced single layer and composite gate dielectrics," *Appl. Phys. Lett.*, vol. 74, p. 2005, Apr. 1999.

[23] T. P. Ma, "Making silicon nitride film a viable gate dielectric," *IEEE Trans. Electron Devices*, vol. 45, p. 680, Mar. 1998.

[24] X. Guo and T. P. Ma, "Tunneling leakage current in oxynitride: dependence on oxygen/nitrogen content," *IEEE. Electron Devices Lett.*, vol. 19, p. 207, June 1998.

[25] H. Y. Yang, H. Niimi, and G. Lucovsky, "Tunneling currents through ultrathin oxide/nitride dual layer gate dielectrics for advanced microelectronic devices," *J. Appl. Phys.*, vol. 83, no. 4, p. 2327, Feb. 1998.

[26] X. Guo, X. Wang, Z. Luo, T. P. Ma, and T. Tamagawa, "High quality ultra thin (1.5nm) TiO_2/Si_3N_4 gate dielectric for deep sub-micron CMOS technology," in *Proc. Int. Electron Devices Meeting, IEDM*, Dec. 1999, p. 137.

[27] H. F. Luan, S.J. Lee, C. H. Lee, S. C. Song, Y. L. Mao, Y. Senzaki, D. Roberts, and D. L. Kwong, "High quality Ta_2O_5 gate dielectrics with $T_{ox,eq} < 10$A," in *Proc. Int. Electron Devices Meeting*, Dec. 1999, p. 141.

[28] D. Park, Q. Lu, T. J. King, C. Hu, A. Kalnitsky, S. P. Tay, and C. C. Cheng, "SiON/Ta_2O_5/TiN gate stack transistor with 1.8nm equivalent SiO_2 thickness," in *Proc. Int. Electron Devices Meeting*, Dec. 1998, p. 381.

[29] G. B. Alers, D. J. Werder, Y. Chabal, H. C. Lu, E. P. Gusev, E. Garfunkel, T. Gustafsson, and R. S. Urdahl, "Intermixing at the tantalum oxide/silicon interface in gate dielectric structures," *Appl. Phys. Lett.*, vol. 73, no. 11, p. 1517, Sept. 1998.

[30] H. Ono and K. Koyanagi, "Formation of silicon–oxide layers at the interface between tantalum oxide and silicon substrate," *Appl. Phys. Lett.*, vol. 75, no. 22, p. 3521, Nov. 1999.

[31] P. K. Roy and I. C. Kizilyalli, "Stacked high-k gate dielectric for gigascale integration of metal–oxide-semiconductor technologies," *Appl. Phys. Lett*, vol. 72, no. 22, p. 2835, June 1998.

[32] B. H. Lee, Y. Jeon, K. Zawadzki, W.-J. Qi, and J. Lee, "Effects of interfacial layer growth on the electrical characteristics of thin titanium oxide films on silicon," *Appl. Phys. Lett.*, vol. 74, no. 21, p. 3143, May 1999.

[33] B. H. Lee, L. Kang, W. J. Qi, R. Nieh, Y. Jeon, K. Onishi, and J. C. Lee, "Ultrathin hafnium oxide with low leakage and excellent reliability for alternative gate dielectric application," in *Proc. Int. Electron Devices Meeting*, Dec. 1999, p. 133.

[34] Y. Ma, Y. Ono, L. Stecker, D. R. Evans, and S. T. Hsu, "Zirconium oxide based gate dielectrics with equivalent oxide thickness of less than 1.0 nm and performance of submicron MOSFET using a nitride gate replacement process," in *Proc. Int. Electron Devices Meeting*, Dec. 1999, p. 149.

[35] E. P. Gusev, M. Copel, E. Cartier, I. J. R. Baumvol, C. Krug, and M. A. Gribelyuk, "High-resolution depth profiling in ultrathin Al_2O_3 films on Si," *Appl. Phys. Lett.*, vol. 76, p. 176, Jan. 2000.

[36] L. Manchanda *et al.*, "Gate quality doped high-k films for CMOS beyond 100nm: 3–10 nm Al_2O_3 with low leakage and low interface states," in *Proc. Int. Electron Devices Meeting*, Dec. 1998, p. 605.

[37] G. D. Wilk, R. M. Wallace, and J. M. Anthony, "Hafnium and zirconium silicates for advanced gate dielectrics," *J. Appl. Phys.*, vol. 87, no. 1, p. 484, Jan. 2000.

[38] R. A. McKee, F. J. Walker, and M. F. Chisholm, "Crystalline oxides on silicon: The first five monolayers," *Phys. Rev. Lett.*, vol. 81, no. 14, p. 3014, Oct. 1998.
[39] W. A. Harrison, *Electronic Structure and the Properties of Solids—The Physics of the Chemical Bond*. San Francisco, CA: Freeman, 1980.
[40] D. Frank *et al.*, "CMOS device and circuit limits," *Proc. IEEE*, vol. 89, Mar. 2001.
[41] K. K. Ng and W. T. Lynch, "Analysis of the gate-voltage dependent series resistance of MOSFETs," *IEEE Trans. Electon Devices*, vol. 33, p. 965, July 1996.
[42] S. Thompson, P. Packan, T. Ghani, M. Stettler, M. Alavi, I. Post, S. Tyagi, S. Ahmed, S. Yang, and M. Bohr, "Source/drain extension scaling for 0.1 μm and below channel length MOSFETs," in *Symp. VLSI Technol. Dig. Tech. Papers*, June 1998, p. 132.
[43] F. A. Trumbore, "Solid solubilities of impurity elements in germanium and silicon," *Bell Syst. Tech. J.*, vol. 39, p. 205, 1960.
[44] P. M. Rousseau, P. B. Griffin, W. T. Fang, and J. D. Plummer, "Arsenic deactivation enhanced diffusion, a time, temperature and concentration study," *J. Appl. Phys.*, vol. 84, no. 7, p. 3593, Oct. 1998.
[45] G. Masetti, M. Severi, and S. Solmi, "Modeling of carrier mobility against carrier concentration in arsenic, phosphorus and boron-doped silicon," *IEEE. Trans. Electron Devices*, vol. 30, p. 764, July 1983.
[46] C. M. Osburn and K. R. Bellur, "Low parasitic resistance contacts for scaled ULSI devices," *Thin Solid Films*, vol. 332, no. 1, p. 428, Nov. 1998.
[47] C. M. Osburn, J. Y. Tsai, and J. Sun, "Metal silicides: Active elements of ULSI contacts," *J. Electron Mater.*, vol. 25, no. 11, p. 1725, Nov. 1996.
[48] P. M. Fahey, P. B. Griffin, and J. D. Plummer, "Point defects and dopant diffusion in silicon," *Rev. Mod. Phys.*, vol. 61, no. 2, pp. 289–384, Apr. 1989.
[49] R. Kasnavi, P. B. Griffin, and J. D. Plummer, "Ultra low energy implant limits on arsenic sheet resistance and junction depth," in *Symp. VLSI Technol. Dig. Tech. Papers*, June 2000, p. 112.
[50] W. Shockley and J. Last, "Statistics of the charge distribution for a localized flaw in a semiconductor," *Phys. Rev.*, vol. 107, no. 2, p. 392, June 1957.
[51] A. Agarwal, D. J. Eaglesham, H. J. Gossmann, L. Pelaz, S. B. Herner, D. C. Jacobson, T. E. Haynes, Y. Erokhin, and R. Simonton, "Boron-enhanced-diffusion of boron: The limiting factor for ultra-shallow junctions," in *Proc. Int. Electron Devices Meeting*, Dec. 1997, p. 467.
[52] A. Agarwal, H. J. Gossmann, D. J. Eaglesham, L. Pelaz, D. C. Jacobson, T. E. Haynes, and Y. Erokhin, "Reduction of transient diffusion from 1–5keV Si^+ ion implantation due to surface annihilation of interstitials," *Appl. Phys. Lett.*, vol. 71, no. 21, p. 3141, Nov. 1997.
[53] J. G. Fossum, K. Kim, and Y. Chong, "Extremely scaled double-gate CMOS performance projections, including GIDL-controlled off-state current," *IEEE Trans. Electron Devices*, vol. 46, p. 2195, NOv. 1999.
[54] C. P. Auth and J. D. Plummer, "Scaling theory for cylindrical, fully depleted surrounding gate MOSFETs," *IEEE Electron Device Lett.*, vol. 18, p. 74, Feb. 1997.
[55] Y. Taur, C. H. Wann, and D. J. Frank, "25 nm CMOS design considerations," in *Proc. Int. Electron Devices Meeting*, Dec. 1998, p. 789.
[56] T. Ohguro, H. Naruse, H. Sugaya, H. Kimijima, E. Morifuji, T. Yoshitomi, T. Morimoto, H. S. Momose, Y. Katsumata, and H. Iwai, "0.12 μm raised gate/source/drain epitaxial channel NMOS technology," in *Proc. Int. Electron Devices Meeting*, Dec. 1998, p. 927.

James D. Plummer (Fellow, IEEE) received the B.S. degree from the University of California, Los Angeles, and the M.S. (EE) and Ph.D. (EE) degrees from Stanford University, Stanford, CA.

He is currently the Frederick E. Terman Professor of Electrical Engineering and Dean of the School of Engineering at Stanford University. He has authored or coauthored over 300 technical papers. His current research interests focus on silicon devices and technology, with a significant part of his work aimed at developing physically based models for silicon structures and fabrication processes. He is also interested in developing new semiconductor devices, including nanoscale structures, high voltage devices, and devices and circuits aimed at special applications.

Dr. Plummer was elected to the National Academy of Engineering in 1996. He received three Best Paper Awards at the International Solid State Circuits Conference in 1970, 1976, and 1978 and the the Solid State Science and Technology Award from the Electrochemical Society in 1991. He has also received several teaching awards at Stanford University.

Peter B. Griffin received the B.E. degree from University College Cork, the M.Eng.Sc. (EE) degree from the National Microelectronics Research Center, Cork, Ireland, and the Ph.D. (EE) degree from Stanford University, Stanford, CA.

He is currently a Research Scientist at Stanford University. He has authored or coauthored more than 50 technical papers, coauthored a new textbook on Silicon VLSI Technology, and consults on advanced process technology with industry. His current research interests are aimed at developing atomistic insights into the fabrication process for silicon devices by a mix of experimental and computational methods.