The Progress and Challenges of Applying High-k/Metal-Gated Devices to Advanced CMOS Technologies

Hsing-Huang Tseng, Ph.D.
Professor of Electrical Engineering
Ingram School of Engineering
Texas State University
601 University Drive, San Marcos, TX 78666,
USA

1. Introduction

1.1 Motivation for implementing high dielectric constant gate dielectric for advanced CMOS scaling

Semiconductor devices need to have good performance, with a low cost and low power dissipation. For decades, research and development of semiconductor processing technology and device integration have focused on enhancing performance and reducing costs using SiO\(_2\) as the gate dielectric and doped polysilicon as the gate electrode. The most effective way to enhance performance and reduce costs is to scale the device gate length and gate oxide. Scaling the gate length results in fabricating more devices per wafer (i.e., increase the device density) and thus reduce the cost per chip, while scaling the gate oxide enhances the drive current and reduces the short channel effects due to gate length scaling. However, as the gate oxide becomes thinner, the power to operate transistors increases because of greater gate oxide leakage current. To resolve this high gate oxide leakage problem, the mechanism of the carriers tunneling through the gate dielectric must be better understood. In an ideal metal-insulator-semiconductor (MIS) device, the current conduction in the insulator should be zero. In a real MIS device, however, current can flow through the insulating film by various conduction mechanisms. The two primary conduction mechanisms for electron tunneling through high quality gate dielectric are discussed below.

1.1.1 Direct tunneling

In a metal-insulator-semiconductor (MIS) stack, when the oxide voltage \(V_{ox}\) is smaller than the metal-insulator barrier height, the electron tunnels directly from the metal electrode into other semiconductor electrode through the insulator. This is known as the direct tunneling process. The equation for direct tunneling current density (current normalized by device area) is proportional to

\[
J_{DT} = A \exp \left( -m \frac{1}{2} V_{b}^{1/2} T_{ox} \right) = B \exp \left( -m \frac{1}{2} V_{b}^{1/2} K \right)
\]

where $m$ is the effective mass of the electron, $V_b$ is the barrier height, $T_{ox}$ is the physical thickness of the gate oxide $\text{SiO}_2$, $K$ is the dielectric constant of the insulator, and $A$ and $B$ are pre-exponential factors. From this equation, one can observe that the $T_{ox}$ or $K$ is the dominating factor in controlling the direct tunnel current density. The tunneling current density increases dramatically as the $\text{SiO}_2$ becomes thinner. In general, the gate leakage increases 100 times for every 0.5 nm that the $\text{SiO}_2$ is thinned. The gate leakage density is as high as the $10^4$ Amp/cm$^2$ range for $\text{SiO}_2$ as thin as 1.1 nm. The high gate leakage increases standby power consumption according to the following equation:

$$P_{\text{STANDBY}} = (I_{\text{subth}} + I_{\text{GIDL}} + I_g) V_{dd}$$ (2)

Where: $I_{\text{subth}}$: subthreshold leakage current
$I_{\text{GIDL}}$: gate-induced drain leakage current
$I_g$: gate leakage
$V_{dd}$: supply voltage

On the other hand, Eq. [1] shows that increasing the dielectric constant of an insulator dramatically reduces the direct tunneling current. This is because the gate oxide physical thickness ($T_{ox}$) and the dielectric constant $K$ of an insulator are correlated by the equivalent oxide thickness (EOT) defined as

$$\text{EOT} = (K_{\text{SiO}_2}/K_{\text{insulator}}) T_{\text{insulator}}$$ (3)

where $K_{\text{SiO}_2}$ and $K_{\text{insulator}}$ are the dielectric constant of $\text{SiO}_2$ and the insulator, respectively, and $T_{\text{insulator}}$ is the physical thickness of the insulator. Based on this definition, an insulator material with a five times greater dielectric constant than $\text{SiO}_2$ would require a five times greater physical thickness than $\text{SiO}_2$ to keep the same EOT as $\text{SiO}_2$. Therefore the tunneling current for a device using this insulator would be orders of magnitude lower than that using $\text{SiO}_2$ because the tunneling leakage current decays exponentially as the insulator becomes thicker.

### 1.1.2 Fowler nordheim tunneling

When the oxide voltage ($V_{ox}$) is greater than the metal/insulator barrier height, the electron tunnels from the metal electrode into the insulator conduction band first and then travels toward the other semiconductor electrode. This is known as the Fowler-Nordheim (FN) tunneling process. The equation for FN tunneling current density is proportional to

$$J_{\text{FN}} = C \exp \left( -m^{1/2} V_b^{3/2} T_{ox} \right) = D \exp \left( -m^{1/2} V_b^{3/2} K \right)$$ (4)

where $C$ and $D$ are pre-exponential factors. From this equation, one can observe that the barrier height is the dominating factor in controlling the FN tunnel current density.

Table 1 compares the exponent (the product of $m$, $V_b$, and $K$) shown in the equations of direct tunneling (low field) and FN tunneling (high field) for insulator materials with different $V_b$ and $K$ values. The results show the following:

1. Although oxynitride processed by incorporating nitrogen into $\text{SiO}_2$ was developed to increase the dielectric constant, the reduction of leakage current density is not enough to be used for highly scaled devices such as for the 45 nm technology node. Even for pure nitride ($\text{Si}_3\text{O}_4$) shown in the table, the exponent shown in the direct tunneling...
The Progress and Challenges of Applying High-k/Metal-Gated Devices to Advanced CMOS Technologies

1.2 Motivation for implementing metal gates for advanced CMOS scaling

1.2.1 SiO\textsubscript{2}/Polysilicon stack

Figure 1 shows a schematic of a MOSFET in which the gate oxide is SiO\textsubscript{2} and the gate electrode is doped polysilicon. Figure 2 shows the equivalent circuit of an MOS capacitor. The total MOS capacitance $C_g$ can be expressed as

$$C_g^{-1} = C_{ox}^{-1} + C_s^{-1} + C_p^{-1}$$

where $C_{ox}$ is the oxide capacitance, $C_s$ is the silicon capacitance, and $C_p$ is the polysilicon gate electrode depletion capacitance.

![Fig. 1.](image1)

![Fig. 2.](image2)

Table 1.

<table>
<thead>
<tr>
<th>Material</th>
<th>$V_b$ (Volts)</th>
<th>$K$</th>
<th>Low Field</th>
<th>High Field</th>
</tr>
</thead>
<tbody>
<tr>
<td>SiO\textsubscript{2}</td>
<td>3.0</td>
<td>3.9</td>
<td>-6.75</td>
<td>-20.3</td>
</tr>
<tr>
<td>Si\textsubscript{3}N\textsubscript{4}</td>
<td>2.0</td>
<td>7.8</td>
<td>-11.0</td>
<td>-22.1</td>
</tr>
<tr>
<td>Ta\textsubscript{2}O\textsubscript{5}, HfO\textsubscript{2}, ZrO\textsubscript{2}</td>
<td>1.5</td>
<td>25</td>
<td>-30.6</td>
<td>-45.9</td>
</tr>
</tbody>
</table>

On the other hand, it is clear that an insulator with a $K$ value of 25 reduces the exponent significantly in the direct tunneling Eq. [1] and more than two times in the FN tunneling Eq. [2].
When the MOSFET is operated in inversion mode, the doped polysilicon gate energy band bending and charge distribution form a thin space-charge region. This results in a finite, bias-dependant value of \( C_p \) and causes polysilicon depletion. The \( C_p \) will reduce the value of the \( C_{ox} \) for an applied gate voltage \( (V_g) \), which, in term, will degrade the MOSFET performance. Although the \( C_p \) can be reduced by increasing the dopant concentration in the polysilicon gate electrode, a high dopant concentration would result in dopant penetrating the gate oxide and induce a threshold voltage \( (V_t) \) instability problem. On the other hand, using a metal gate as the gate electrode could eliminate the polysilicon gate depletion problem without \( V_t \) instability concern because there is no need for dopant to be incorporated into the gate electrode. Another advantage of a metal gate is that the resistance of the metal gate electrode is less than a polysilicon gate.

1.2.2 **High dielectric constant insulator compatibility with gate electrode**
The compatibility of polysilicon gate electrodes with high dielectric constant (high-k) insulators raises some concern. Most metal oxides with a high dielectric constant used as a gate insulator react with polysilicon and degrade the gate dielectric. Furthermore, this interaction makes it difficult to control the MOSFET threshold voltage. On the other hand, a metal gate is more compatible with high-k metal oxides.

2. Materials screening of high dielectric constant insulators and metal gates

2.1 **High-k gate dielectric screening**

2.1.1 **Issues of high-k gate dielectric**
Some fundamental issues with implementing a high-k gate dielectric in MOSFETs are as follows:

- Thermodynamic stability of high-k on silicon
- Trade-off between the dielectric constant \( (K) \) and band gap \( (E_g) \)
- Film microstructure: crystalline vs. amorphous
- \( O_2 \) and dopant diffusion through the grain boundary
- Impact of the amorphous interfacial layer (IL) on the overall dielectric constant of the film, EOT scalability, interfacial roughness, and MOSFET mobility
- Possible mobility degradation and high fixed charge caused by a high-k insulator
- Compatibility with the gate electrode
  a. Thermodynamic stability of high-k on silicon
     An analysis of the Gibbs free energies governing the following chemical reactions for metal-Si-oxygen ternary systems is important in predicting stability. To avoid instability with Si to form SiO₂,
     \[
     Si + M_{ox} \rightarrow M + SiO_2
     \]
     To avoid silicide formation,
     \[
     Si + M_{ox} \rightarrow M_Si + SiO_2
     \]
  b. Trade-off between the Dielectric Constant \( (K) \) and Band Gap \( (E_g) \)
     From the direct tunneling Eq. [1], it is desirable to find an insulator with a high dielectric constant and high barrier to ensure low gate leakage current density. Since the barrier height values for most high-k metal oxides are not reported, the closest and most readily available indicator for the band offset is the band gap values. Figure 3 shows the plot of band gap versus dielectric constant for various metal oxides.
The Progress and Challenges of Applying High-k/Metal-Gated Devices to Advanced CMOS Technologies

Fig. 3. The trade-off between dielectric constant and band gap limits the choice of metal oxides.

c. Film Microstructure: Crystalline versus Amorphous
   For polycrystalline metal oxide, the triple point may generate defects/voids, which will cause a device yield issue. In addition, oxygen, dopant, and impurities diffuse swiftly in the polycrystalline structure primarily through the grain boundary and degrade the electrical properties of the gate stack. Another potential concern is controlling the grain size among small devices and wafers. Amorphous metal oxides can reduce O\textsubscript{2} and dopant diffusion and lower defectivity; however, they usually have a lower dielectric constant than those metal oxides with a polycrystalline structure.

d. Oxygen Diffusion through the Grain Boundary of Metal Oxide
   There is a distinct processing difference between the metal oxides and conventional thermal oxide SiO\textsubscript{2}. Metal oxides are deposited on the silicon substrate instead of thermally grown like SiO\textsubscript{2}. The intrinsic quality of the deposited film is inferior to thermally grown film. A post-deposition anneal under dilute oxygen ambient is necessary for high performance devices. Most metal oxides with a high dielectric constant, however, form a crystalline structure after a relatively high temperature anneal. Therefore the oxygen contained in the ambient of the post-metal oxide deposition anneal diffuses through the grain boundaries of the metal oxides and reacts with silicon substrate, which forms a SiO\textsubscript{2} interfacial layer (IL).

e. Impact of the Amorphous Interfacial Layer (IL) on the Overall Dielectric Constant of the Insulator, EOT Scalability, Interfacial Roughness, and MOSFET Mobility
   The SiO\textsubscript{2}-like interfacial layer reduces the overall dielectric constant of the bi-layer gate dielectric (high-k on top of a SiO\textsubscript{2} IL). The IL is about 1 nm thick, which would make scaling the EOT to less than 1 nm difficult. The interface between the IL and silicon substrate is rougher than the interface between conventional thermally grown SiO\textsubscript{2} and...
silicon, which may degrade channel carrier mobility and generate interface state defects. Figure 4 shows a transmission electron microscopy (TEM) image of the metal oxide MOS structure and the key concerns about the gate stack.

Fig. 4.

f. Possible Mobility Degradation and High Fixed Charge Caused by the High-k Insulator

Soft phonons in the metal oxide bonding structure contribute to the high dielectric constant property of metal oxides. Soft phonons are generated by high atomic number atoms resonating in their bonding structures. These phonons make a lattice contribution to the overall polarizability and therefore the high dielectric constant. There is a concern that the mobility of the channel carriers may be degraded by interactions with these soft phonons.

2.1.2 High-k Gate Dielectric Candidates
a. Metal Oxide

After eliminating the metal oxides that are thermodynamically reactive with silicon substrates or that have a relatively low dielectric constant or small band gap, the remaining candidates fall under group IVB, IIIA, and IIIB of the periodic table.

i. Metal Oxides Used for Memory Capacitors [1, 2, 3, 4]

Candidates such as TiO$_2$ and Ta$_2$O$_5$ have the advantage of having a relatively high dielectric constant and a history of processing in the industry. However, the following concerns make them unattractive for logic devices:

- Small band gap
- Instability with silicon substrates
- High density of oxygen vacancies
- Require an oxygen anneal to improve film quality, which results in oxidation of bottom leading to an undesirable increase in EOT
• Unstable microstructure

ii. Group IIIA and IIIB Metal Oxides: Al₂O₃ and La₂O₃ [5, 6, 7]
Both Al₂O₃ and La₂O₃ are thermodynamically stable with silicon substrates. In addition, Al₂O₃ is amorphous at 1000°C and has a relatively high band gap (8.7 eV ~ SiO₂’s value). However, it has a relatively low dielectric constant; high oxygen, B, and P diffusion; and easily absorbs H₂O. La₂O₃ has a relatively high dielectric constant (K ~ 27), but the band gap is small (4.3 eV) and it very easily absorbs moisture from the ambient.

iii. Group IVB Metal Oxides: HfO₂ and ZrO₂ [8-18]
These metal oxides have reasonably high dielectric constants and band gaps (see Figure 3). Both ZrO₂ and HfO₂ devices have demonstrated a many orders of magnitude reduction in gate leakage with an EOT around 1.0 nm and well-behaved transistors. However, they have high O₂ and dopant diffusivity due to their crystalline microstructure. ZrO₂ has a relatively stronger interaction with polysilicon gates than HfO₂. Figure 6 shows the interfacial layer thickness beneath ZrO₂ increases as the post-deposition anneal temperature increases from 550 to 650°C.

![Fig. 6. ZrO₂ Interfacial layer](image)

Figure 7 shows the x-ray photoelectron spectroscopy (XPS) spectra of the interface between ZrO₂ and the polysilicon gate. The polysilicon gate was deposited in situ on ZrO₂. XPS reveals that ZrO₂ decomposes into a Zr metal compound when the gate stack is annealed in nitrogen at 950°C under ultra-high vacuum leading to a high gate leakage current. It also shows the formation of interfacial SiO₂ between ZrO₂ and the polysilicon gate during polysilicon deposition. These results suggest a strong interaction between ZrO₂ and SiH₄ during polysilicon deposition at 550 to 620°C.

On the other hand, HfO₂ metal oxide is thermodynamically more stable with silicon than ZrO₂. Figure 8 shows the XPS spectra of the interface between HfO₂ and the polysilicon gate. Unlike the ZrO₂ film, the HfO₂ film remains stable after polysilicon deposition and a post-anneal in nitrogen up to 950°C.
Adding silicon to metal oxide can maintain the amorphous phase up to a medium temperature such as 800°C depending on the silicon concentration. These metal silicates are thermodynamically stable with the silicon substrate. Figure 9 shows the TEM cross-sections of a gate stack composed of $\text{ZrSi}_x\text{O}_y$ silicate deposited on a silicon substrate with an aluminum metal electrode. No interfacial layer forms between the zirconium silicate and the silicon substrate after annealing at 800°C for 30 minutes in nitrogen ambient. The interface is atomically sharp, and the film remains amorphous after the anneal. However, there is a possible phase separation with a high temperature anneal and the dielectric constants are lower than metal oxide candidates such as $\text{ZrO}_2$ and $\text{HfO}_2$. 
2.1.3 High-k dielectric deposition techniques

An important factor in determining the final choice of high-k dielectric is the deposition process, which must be compatible with current CMOS processing, cost, and throughput. In general, there are four major deposition techniques:

- **Metal organic chemical vapor deposition (MOCVD)** is commonly used, but the C, H, and OH impurities contained in the film are a major concern.

- **Physical vapor deposition (PVD)** is good for evaluating new materials. However, the purity of the target and plasma-induced damage are common concerns.

- **Molecular beam epitaxy (MBE)** is good for interfacial control, but the throughput is rather low.

- **Atomic layer deposition (ALD)** has high uniformity control and good conformality. However, throughput is low and the process is sensitive to surface preparation.
Potential contamination with Cl, C, H, and OH impurities is also a concern. Figure 10 shows a typical ALD process to fabricate metal oxide [23, 24]. In summary, same high-k materials fabricated by different deposition tools, processes, and precursors result in different properties. The final choice of the deposition technique needs to balance cost, throughput, tool reliability, film properties, and device performance and reliability.

2.1.4 High-k gate dielectric device integration issues [17]

a. Device Size Dependence of Gate Current Density

Figure 11 shows that the gate current density of the ZrO$_2$/polysilicon gate stack has a strong dependence on device size. At 2 V, the leakage current density for the 14 µm NMOS (PMOS) device is 9X (7X) that of a 1.4 µm device. Figure 12 shows the TEM cross-section of the PMOS capacitor with a longer gate length. It is clear that some reactions have occurred at the polysilicon/ZrO$_2$ interface. The TEM cross-section (Figure 13) shows a nodule extending above and below the polysilicon/ZrO$_2$ interface. The penetration into ZrO$_2$ creates a conduction path that results in a high gate leakage current. The longer gate length results in a higher probability that conduction paths will be formed.
Fig. 13.

b. Lateral Oxidation Model [17]

Figure 14 shows that no nodule is observed at the edge of the devices. A detailed XTEM analysis found the first nodules located ~ 2-3 µm from the edge of the polysilicon gate. This observation suggests that devices with 4-6 µm or smaller gate lengths are nodule-free. It is proposed that active oxygen diffuses through the metal oxide and grows an oxide (SiO₂) at the polysilicon interface (Figure 15). The oxide prevents nodule formation during a high temperature anneal such as a source/drain anneal. Oxidation at the center of large devices is insufficient to prevent the formation of nodules, which cause high leakage current.

![PMOS Capacitor. 1µm from Edge](image)

Fig. 14.

![Diagram of Oxidation Model](image)

Fig. 15.
2.1.5 High-k dielectric summary
Select high-k metal oxides with thin EOTs have demonstrated an orders of magnitude lower gate leakage than SiO$_2$. However, it is still a challenge to achieve an EOT thinner than 1.0 nm after transistor fabrication. Degradation in device mobility is observed when using a high-k dielectric. This is more problematic for low EOT applications. Since compatibility with polysilicon gate electrodes is a major concern for some potential metal oxides and metal gate has several advantages, a high-k/metal gate stack is the choice for advanced devices.

2.2 Metal gate electrode materials screening
2.2.1 Materials constraints
A key issue for metal gate materials research is controlling the work function of metal gate electrodes after CMOS processing. There are two choices of metal gate implementation. The first type is a single metal electrode with a work function near the mid-gap (~4.6 eV) using additional processing or incorporating additional materials to control the work function of CMOS devices. The second type is a dual metal gate electrode with one metal having a work function (4.1 eV) near the conduction band of the silicon substrate ($E_C$) for NMOS and the other one having work function (~5.2 eV) near the valence band of the silicon substrate ($E_V$) for PMOS (Figure 16). The metal electrode materials should have thermal, chemical, and mechanical stability with the high-k gate dielectric and surrounding material during CMOS processing. It is desirable to have a metal gate with a low sheet resistance that is compatible with CMOS process integration, either a conventional “gate first” or replacement “gate last” approach. Metallic elements, compounds (nitrides, silicides, carbides, borides, etc.), and solid solutions are possible candidates.

![Fig. 16.](image)

Work functions can be obtained from MOSCAPs of various oxide thicknesses using the following equation:

$$V_{fb} = \phi_M + Q_F/C_{OX}$$

An intersect in the Y-axis of the $V_{fb}$ versus EOT plot is the work function (Figure 17). Table 2 shows the work function values for some potential metal gate electrode candidates.
The Progress and Challenges of Applying High-k/Metal-Gated Devices to Advanced CMOS Technologies

Fig. 17. [25]

![Graph showing flatband voltage vs oxide thickness with various materials and work functions.]

Table 2.

<table>
<thead>
<tr>
<th>Gate Material</th>
<th>Work Function (MOS) (eV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ti</td>
<td>4.17, 4.33, 4.6</td>
</tr>
<tr>
<td>TiN</td>
<td>4.95</td>
</tr>
<tr>
<td>TiSi2</td>
<td>3.67-4.25</td>
</tr>
<tr>
<td>Zr</td>
<td>4.05</td>
</tr>
<tr>
<td>ZrN</td>
<td>4.6</td>
</tr>
<tr>
<td>ZrSi2</td>
<td>-</td>
</tr>
<tr>
<td>Ta</td>
<td>4.25, 4.6, 4.15-4.25</td>
</tr>
<tr>
<td>TaN</td>
<td>5.41</td>
</tr>
<tr>
<td>TaSi2</td>
<td>4.15</td>
</tr>
<tr>
<td>Nb</td>
<td>4.3, 4.02-4.3</td>
</tr>
<tr>
<td>NbN</td>
<td>-</td>
</tr>
<tr>
<td>NbSi2</td>
<td>4.35-4.53</td>
</tr>
<tr>
<td>W</td>
<td>4.75, 4.72, 4.55-4.63</td>
</tr>
<tr>
<td>WNx</td>
<td>5</td>
</tr>
<tr>
<td>WSi2</td>
<td>4.55-4.8</td>
</tr>
<tr>
<td>Mo</td>
<td>4.64, 4.53-4.6</td>
</tr>
<tr>
<td>MoN</td>
<td>5.33</td>
</tr>
<tr>
<td>MoSi2</td>
<td>4.6-4.8, 4.9</td>
</tr>
<tr>
<td>Al</td>
<td>4.1</td>
</tr>
</tbody>
</table>

www.intechopen.com
2.2.2 Metal electrode material deposition method and film properties

Metal deposition processing parameters and post-deposition processing affect the metal film properties such as resistivity, microstructure (grain size and orientation), stress, and adhesion. Deposition processes such as CVD and PVD are common methods. In general, CVD films provide better conformality and negligible damage compared to PVD films, but the process may incorporate contaminants. Atomic layer deposition (ALD) shows excellent conformality, but throughput is low. Figures 18 and 19 show the impact of the metal gate deposition process on device performance and gate leakage current, respectively. The device with a SiO₂/PVD TiN metal gate stack results in lower mobility than CVD TiN or polysilicon gated devices. The devices with a SiO₂/PVD TiN gate stack result in a 100X higher gate leakage current than devices fabricated with CVD and ALD.

Fig. 18. [26]

Fig. 19. [27]
2.2.3 Metal gate device integration issues
The dual metal gate approach that needs different metal gate materials for NMOS and PMOS, respectively, increases process complexity dramatically. Contamination (mostly from CVD) and plasma damage (mostly from PVD) affect gate device parameters such as \(Q_f\), \(D_{it}\), \(V_t\) stability, and gate oxide integrity. Another concern with metal gates is poor oxidation resistance. The etchability of metal materials, especially selectivity to metal oxides, is a key area for process development.

a. Conventional “Gate First” CMOS Integration
Conventional gate first integration involves the ability to etch the gate material. For example, Figure 20 shows a nitride/W/TiN gate electrode stack on top of gate oxide. The gate electrode materials must be protected from oxidation or attack by wet chemicals. The gate electrode materials must also be stable with their surrounding materials during high temperature steps. These requirements may limit the choices of materials for metal gate materials along with alternative high-k dielectrics.

![Fig. 20. [28, 29]](image)

b. Replacement “Gate Last” CMOS Integration
This integration eliminates many constraints posed by conventional “gate first” integration such as process-induced damage, the requirement to etch new materials, thermal/chemical stability concerns, stress-induced diffusion issues. After completing conventional MOSFET fabrication with replacement polysilicon, gate oxide is deposited and using chemical-mechanical polishing (CMP) technique to flatten the top surface (Figure 20a). A wet etch or dry etch process is used to etch off the polysilicon gate (Figure 20b) followed by a new metal gate electrode material and a new gate dielectric deposition (Figure 20c). Another CMP process or a second patterning of the gate is used to form the final structure.
c. Summary

A single metal gate with a mid-gap work function may not be able to achieve a low MOSFET threshold voltage and boost performance; however, it has potential for fully depleted silicon-on-insulator (FDSOI) applications. Dual metals with work functions similar to n⁺ polysilicon and p⁺ polysilicon pose significant integration challenges. Thermodynamic and mechanical stability is an important issue in the choice of metal gate materials. Both the gate first and gate last integration have advantages and disadvantages. The final choice of integration scheme will be decided by performance, yield, and cost.

3. Device characteristics using High-k/Metal Gate (HKMG) stack

3.1 Impact of defects in the HKMG stack on device performance and reliability

As mentioned in section 2.1.1.d, the deposited high-k gate dielectric contains a high defect density in the bulk even after a post-deposition anneal (PDA). The quality of the interface between the silicon substrate and interfacial layer (IL) is not as good as the interface between silicon and conventional thermally grown gate oxide SiO₂. A TEM cross-section of the HKMG gate stack, shown in Figure 21, highlights the different regions of the gate stack. The defects in the bulk of the high-k gate dielectric and at the interface between the silicon substrate and IL have a significant impact on device performance and reliability, such as threshold voltage stability. As discussed in Section 2.1.2.a.iii, the best candidate for a high-k dielectric is Hf-based metal oxide. Therefore the following discussion is based on HfO₂/metal gate stacks.
3.1.1 Si/IL interface improvement – stressed relaxed pre-oxide [31]

Depositing high-k on top of thick, high quality, thermally grown SiO₂, taking the advantage of its better interface quality, is not desirable because we need a thin gate dielectric to increase transistor speed. To solve this problem, a stress relieved pre-oxide (SRPO) process has been developed to improve the interfacial properties between the high-k dielectric and silicon substrate while maintaining the required thinness to meet the speed enhancement requirement for integrated circuits. The experiment discussed here is as follows. The SRPO is formed by growing a relatively thick thermal oxide with a high temperature anneal (higher than the SiO₂ glass flow temperature of ~ 980°C) for stress relief followed by etching the thermal oxide back to 10 Å using a diluted 700:1 hydrofluoric acid: H₂O solution. The HfO₂ is then deposited by ALD. After the high-k dielectric undergoes a PDA, a TaSiN metal gate is deposited. A commercial CMOS process technology with a source/drain anneal at 1000°C was used to fabricate metal gate/high-k stack nMOSFETs on bulk silicon [32]. Fig. 22 compares the threshold voltage (V_t) shift under constant voltage stress for the control split using the standard process (an RCA clean followed by ALD HfO₂ with a TaSiN metal gate) and the new SRPO process with a TaSiN/HfO₂ gate stack for short channel devices (W/L = 10 µm/0.15 µm). The SRPO with a TaSiN/HfO₂ stack results in a 3X smaller V_t shift than the standard process. These results suggest that devices with the standard process suffer process-induced gate edge damage during transistor fabrication, which increases the V_t shift due to greater trap generation. The process-induced gate edge damage is reduced significantly when using SRPO due to the high quality interfacial layer under the HfO₂, which suppresses interface trap and border trap generation during constant voltage stress. To assess process-induced gate edge damage, the charge pumping current was measured on short channel devices using a pulse string with a fixed base level and varying pulse heights [33] under drain or source bias to detect local charge at the gate edge. The results show less interface and border state trap density for HfO₂/SRPO devices than for HfO₂/RCA devices. The stressed charge pumping results clearly show that the SRPO pre-treatment is much more robust than the RCA pre-treatment under process-induced local charge generation near gate edges. Fig. 23 shows that the normalized transconductance (G_m) of TaSiN-gated short channel devices with SRPO is higher than the standard pre-treatment due to better interface properties with the SRPO process. The fundamental difference between a chemical
oxide formed after an RCA clean and SRPO is as follows. The quality of the chemical oxide is poor and has a lower density than thermally grown SiO\(_2\). It is susceptible to process-induced gate damage. On the other hand, for thermally grown SiO\(_2\), a mechanical stress builds up as the SiO\(_2\) becomes thicker during oxidation caused by the molar volume mismatch between Si and SiO\(_2\) and their different expansion coefficients. The stress degrades the Si-O bonding configuration near the interface between SiO\(_2\) and the silicon substrate. The quality of the sub-oxide layer underneath the bulk SiO\(_2\) is poor. It is therefore important to relieve the stress build-up during oxidation to have a high quality sub-oxide. Annealing the oxide at a temperature higher than the SiO\(_2\) glass flow temperature (~980°C) allows the SiO\(_2\) to “flow” and the bonding configuration near the interface to be rearranged, improving the sub-oxide quality. The SRPO process grows a relatively thick thermal oxide during a high temperature (~980°C) anneal, resulting in a high quality sub-oxide followed by an etch back to 10 Å using a diluted HF solution. The 10 Å SiO\(_2\) serving as a pre-oxide before high-k deposition thus becomes a high quality film. Excellent cross-wafer inversion \(T_{ox}\) uniformity is demonstrated using the SRPO pretreatment for an array of twenty-eight 10 \(\mu\)m \(\times\) 10 \(\mu\)m (W/L) devices as shown in Fig. 24.

---

![Graph](image1)

**Fig. 22.**

![Graph](image2)

**Fig. 23.**
The other approach to improve interface robustness is to incorporate deuterium into the interface between the silicon substrate and IL because the Si-D bonds are much harder to break than Si-H bonds. One effective way to incorporate deuterium is to use D₂O instead of H₂O as the oxidant precursors during ALD [34]. Introducing deuterium in situ during ALD is an effective way to passivate the Si dangling bonds because it does not require the pre-existing hydrogen to be replaced. Hydrogen can be introduced easily from the processes after high-k gate dielectric deposition. Fig. 25 compares the threshold shift for nMOSFETs under positive bias stressing at 25°C and 125°C. A significant reduction in $V_t$ shift is observed for D₂O-processed HfO₂ devices. The presence of deuterium at the interface is supported by the low energy secondary ion mass spectroscopy (SIMS) analysis (Fig. 26), which reveals a spike of deuterium at the 7E17 at/cm³ level. To improve the depth resolution of SIMS, we used 100 Å D₂O-processed HfO₂ to prepare the test wafers. The SIMS analysis of another sample prepared by using 100 Å H₂O-processed HfO₂ shows no deuterium at the bottom interface.
3.1.3 Fluorine passivation coupled with SRPO [35]

In addition to the defects at the interface, defects contained in the bulk HfO$_2$ will degrade device reliability and performance. Incorporating fluorine into the bulk HfO$_2$ and interfaces can passivate the defects, thus improving the device robustness and speed. Combining SRPO interface engineering and defect passivation with fluorine in a high-k/Ta$_x$C$_y$ stack showed excellent V$_t$ stability [35]. The positive bias temperature instability (PBTI) time to failure (TTF)-lifetime extraction using stress voltages in the direct tunneling regime is shown in Fig. 27. Fluorinated devices exceed the V$_t$ TTF lifetime target (<30 mV V$_t$ shift in 10 years at 105°C for V$_{dd} = 1$ V) with a sufficient margin and reveal about a four orders of magnitude longer PBTI lifetime than the control that does not incorporate fluorine. SIMS analysis results confirmed the incorporation of fluorine at the interfaces and bulk high-k [35], which is expected to reduce the defect density and thus reduce charge trapping. It has been shown that oxygen vacancies are the major defects in the bulk HfO$_2$ [36]. Table 3 shows the results of atomistic calculations of the formation energies of fluorine-passivated oxygen vacancies in the bulk of the Hf-based high-k using density functional theory (DFT) [37, 38] as implemented in the local orbital SIESTA code [39]. The results show that the fluorine passivation of threefold (V3) or fourfold (V4) oxygen vacancies is energetically favorable and can therefore lead to less trapping. Fig. 28 shows the structure used to calculate a fluorine atom at a V3 site. The structure of the fluorine-passivated V4 defect (not shown) is similar. To electrically support the bulk trap density reduction from fluorine incorporation, stress-induced leakage current (SILC) was measured. Fig. 29 compares the SILC of fluorinated devices with the control. Fluorine incorporated in the bulk reduces the SILC significantly. The results are consistent with bulk defect passivation by incorporating fluorine. Finally, Fig. 30 shows the gate leakage for the fluorinated high-k/Ta$_x$C$_y$ stack devices is over 4 orders of magnitude lower than that of silicon oxynitride. The CETinv is the sum of EOT and the quantum mechanical effect in the silicon substrate, which is around 0.4 nm.
Table 3.

<table>
<thead>
<tr>
<th>Reaction</th>
<th>Energy Gain (eV)</th>
<th>F Coordination #</th>
<th>F Charge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hf-V3-Hf + F(interstitial) → Hf-F-Hf</td>
<td>-7.66</td>
<td>3</td>
<td>7.18</td>
</tr>
<tr>
<td>Hf-V4-Hf + F(interstitial) → Hf-F-Hf</td>
<td>-7.66</td>
<td>3–4</td>
<td>7.18</td>
</tr>
</tbody>
</table>

Fig. 27.

Fig. 28.

Fig. 29.
Fig. 30.

3.2 Threshold voltage turning

3.2.1 Effective work function of metal gates

In an nMOSFET device, threshold voltage ($V_t$) is the voltage at which electrons in the inversion layer formed at the substrate Si/dielectric interface are sufficient to produce a conducting path between the MOSFET source and drain (S/D). In CMOS applications, the effective work function (EWF) of metal gate electrodes is an important parameter as it determines the flatband voltage ($V_{fb}$) and, subsequently, the $V_t$ of MOSFETs. If no charge is present in the oxide or at the Si/dielectric interface, the $V_{fb}$ equals the work function difference between the gate metal and the semiconductor substrate as shown in Eq. [7]. The work function values shown in Table 2 are based on this simple method. However, the work function of the metal, and thus the $V_t$ obtained after the CMOS processing, is likely different from that obtained from Eq. [7]. The work function obtained after the CMOS processing is called the EWF, which determines the final $V_t$ of CMOS.

A precise measurement of the metal gate EWF is crucial to identifying the optimal electrode material. However, EWF measurements of metal electrodes on high-k dielectrics have shown a dependence on the particular dielectric material and are further complicated by Fermi-level pinning at the high-k/metal interface [40]. Insufficient understanding of metal-electrode systems and their interactions with underlying dielectrics has contributed to inconsistent EWF values. It has further been reported that factors such as specific processing and associated thermal budgets affect the final EWF of high-k/metal electrode systems [41] due to composition variations and the temperature-driven crystalline phase production of metal (metal oxides, metal nitrides, and metal oxynitrides) electrodes.

The EWF of metal electrodes on high-k material may be extracted more precisely from MOS capacitors fabricated by depositing and patterning the high-k dielectric and metal gate capacitor structures on a “terraced” oxide layer consisting of incrementally thicker layers of thermally grown SiO$_2$ [42]. The multiple oxide thicknesses (1.0, 1.5, 2.5, 3.5 nm) on a single wafer are achieved by growing a relatively thick oxide and selectively etching regions with diluted hydrofluoric acid. Complete details of this etch process and evaluated work function results have been described previously [42]. Fig. 31 illustrates the bottom SiO$_2$ terrace step thickness measurements recorded by spectroscopic ellipsometry (SE) in a diameter scan across the wafer. The wafer image insert in Fig 30 illustrates the concentric thickness bands. To evaluate metal gate work functions on high-k films, a fixed amount of high-k (~ 2.0 nm) is deposited on the terraced oxide followed by ALD of a 10 nm TiN (or other metal) gate...
The Progress and Challenges of Applying High-k/Metal-Gated Devices to Advanced CMOS Technologies

The EWFs for terraced oxide stacks were extracted using Eq. [7b] [42], which is a simplified form of the general model given by R. Jha [43] (Eq. [7a]) to enable linear extraction of the $V_{fb}$-EOT relationship:

$$V_{fb} = \Phi_{ms} - \frac{Q_f * EOT}{\varepsilon_{ox}} - \left( \frac{Q_i * t_h}{\varepsilon_h} \right) - \frac{1}{\varepsilon_{ox}} \int_{0}^{t_h} \int_{0}^{t_h} x \rho_b(x) dx - \frac{1}{\varepsilon_{ox}} \int_{t_h}^{t - t_h} x \rho_{ox}(x) dx \right), t_h = \frac{\varepsilon_h}{\varepsilon_{ox}} EOT_h$$

$$V_{fb} = \frac{\Phi_{ms} - \frac{Q_i * EOT_h}{\varepsilon_{ox}} - \frac{\rho_b * \left( \frac{\varepsilon_h}{\varepsilon_{ox}} \right) * EOT_h^2}{2 * \varepsilon_{ox}} - \frac{Q_f * EOT}{\varepsilon_{ox}}}{\varepsilon_{ox}}$$

With the terraced oxide structure, a constant fixed interface charge between the Si substrate and dielectric ($Q_i$) is maintained across the varying oxide thicknesses, minimizing wafer-to-wafer variation associated with multiple wafer extraction methods. $Q_i$ can be calculated from the slope of the $V_{fb}$-EOT relationship. The SiO$_2$ bulk charge term in Ref. [43] can be neglected to simplify the extraction since this bulk charge contribution is far less than that of $Q_i$ [44]. The extracted y-axis ordinate intercept value contains the contribution of the metal WF as well as the high-k/SiO$_2$ interface charge ($Q_i$) and high-k bulk charge density ($\rho_b$) terms ($t_h$: high-k physical thickness, $t_{ox}$: SiO$_2$ physical thicknesses, $t$: total physical thickness of high-k/SiO$_2$ stack; EOT$_h$: high-k EOT; EOT: EOT of high-k/SiO$_2$ stack, $\varepsilon_h$: high-k dielectric constant, $\varepsilon_{ox}$: SiO$_2$ dielectric constant). The contributions of charges in the high-k on $V_{fb}$ are controlled and can be minimized (~ +50 mV) by using a fixed and thinned high-k film (2–3 nm), thus enabling accurate extraction of the EWF. The resultant terraced oxide capacitors exhibit excellent linear $V_{fb}$-EOT fits with minimal effects from variations in fixed charge at the Si/SiO$_2$ interface.

### 3.2.2 Dielectric capping for work function tuning

Although TaSiN (nMOS), TiN (pMOS), and Ru (pMOS) on HfO$_2$ have been demonstrated in CMOS integration [45, 46], the use of dual metal gates must address the significant complexity of optimizing two different metal etch processes. A single metal gate approach for CMOS integration provides several advantages over the dual metal electrode process, specifically a more straightforward integration and less demanding gate etching process...
optimization and control. A single metal gate, however, requires tuning the EWF value to obtain the proper n- and p-type threshold voltage. These tuning efforts have given rise to an extensive study of the impact of capping layer, a thin metal oxide incorporated between the Hf-based high-k dielectrics and metal gates, whereby a single metal gate CMOS process can be implemented with either single or dual cap layer approaches [47]–[52].

Thin cap layers (≤ 1 nm) such as Dy2O3 and Al2O3 deposited on the high-k gate dielectric followed by thermal annealing drives the metal atoms of the cap layer into the high-k layer (and bottom SiOx interface layer). Appropriate control of cap layer doping (depth of diffusion) has tuned the EWF of metal electrodes by dipole formation at the interface between the high-k and bottom interfacial SiOx layer. The EWF of metal electrodes can be shifted toward the valence and conduction bands of Si using AlOx and LaOx capping layers, respectively [52]–[55]. Performance data suggests that AlOx capping is also effective after high temperature processing, indicating that Al atoms need to diffuse to the interface between the HfO2 and the interfacial SiO2 layer to shift the EWF [56]. Modulating LaOx within the HfO2 has been shown to contribute to this EWF shift as a result of La concentration at the SiOx/HfO2 interface. The relative concentration trends with the amount of EWF shift [57]. Coincident Hf and La EELS element profiles and SIMS profiles in Fig. 32 and Fig. 33, respectively, verify that La migrates from a La2O3 cap layer into un-annealed HfO2 [58]. These observations can make the dipole moment formed at the internal interface a more feasible explanation for EWF tuning.

Fig. 32. EELS scan showing intermixing of HfSiO and La2O3.

Analysis of the energy band diagram of the multilayer dielectric stack suggests that the EWF of the gate stack can be modified by the dipole at the interface between two dielectric layers, typical of the high-k gate stack (i.e., the Si/SiOx/HfO2/electrode, where a SiO2-like IL is formed). As indicated above, such a dipole layer can be formed by introducing metal ions into the IL near its interface with the high-k film. The electronegativity of the metallic elements with respect to Hf atoms presents a plausible case for a model explaining EWF tuning [59]. Fig. 34 illustrates the effect of a band offset change between the high-k and interfacial oxide induced by incorporating metal ions into the interfacial layer. A metallic element can be incorporated in the IL by diffusion during thermal processing, specifically
The Progress and Challenges of Applying High-k/Metal-Gated Devices to Advanced CMOS Technologies

with a high temperature (~ 1000°C) S/D dopant activation anneal. Based on their electronegativity relative to Hf, metallic elements can generate either “positive” (i.e., Al (Ti)-O-Hf) or “negative” dipoles resulting in a higher or lower EWF, respectively. Furthermore, ab initio calculations show that the Al-O-Hf dipole results from substituting Al for Si in SiO$_2$ near its interface with HfO$_2$ thereby significantly reducing the SiO$_2$/high-k valence band offset and thus effectively increasing the EWF [60].

![Fig. 33. SIMS profile confirms intermixing of La$_2$O$_3$ with the HfSiO layer.](image)

![Fig. 34. Band diagram illustrating the EWF change of the metal gate depending on the type of dipole.](image)

In nFETs, La is found to be the most effective dopant based on its overall effect on $V_t$, EOT scaling, mobility, and reliability [50, 53, 54]. An Al$_2$O$_3$ cap has been widely used for pFETs, but it increased the EOT [52] (since it has a relatively lower dielectric constant value) as well as raises reliability concerns due to Al diffusing into the interfacial oxide layer [52]. In addition, the $V_{fb}$-EOT roll-off phenomenon (see next section) was observed in Al$_2$O$_3$-capped gate stacks, making it even more difficult to achieve a proper $V_{fb}$ ($V_t$) for pMOS.

### 3.2.3 $V_{fb}$-EOT Roll-Off

When gate stack film systems consisting of a metal electrode, high-k dielectric, and SiO$_2$ IL were used in devices of practical interest with scaled EOTs, their $V_{fb}$ values at thin gate
stacks were found to be significantly less than those obtained in test structures with thicker gate stacks. This is most remarkable in gate stacks with high EWFs [61], as seen in Fig. 35 showing a grand plot of $V_{fb}$ vs. EOT in pMOSCAPs with various metal gates fabricated on terraced oxide structures. The $V_{fb}$ roll-off starts at a certain minimal thickness of the SiO$_2$ IL and increases as the SiO$_2$ becomes thinner. One can clearly see a gradual reduction of $V_{fb}$ as the gate stack EOT scales below a certain value. It is more prominent in gate stacks annealed at higher temperatures and/or for a longer anneal time. This $V_{fb}$-EOT roll-off phenomenon is a thermally activated process that suggests an intrinsic relation to the EWF values of the gate stack and/or the change in electrical integrity of the physically scaled bottom interface layer.

![Fig. 35. Dependence of $V_{fb}$ on EOT in high-k terraced oxide capacitors with metal electrodes of different WF values.](image1)

![Fig. 36. Advantages of fluorine incorporation for $V_{fb}$-EOT roll-off reduction](image2)

The likely root cause of the $V_{fb}$-EOT roll-off problem in HKMG devices is the oxygen vacancies in the high-k stack, which trigger the generation of defects at the bottom of the SiO$_2$/Si interface. Therefore processes that can minimize the oxygen vacancy density in the high-k bulk and/or enhance the robustness of the SiO$_2$/Si interface are critical to reducing the $V_{fb}$-EOT roll-off in advanced HKMG devices. From the point of view of interface quality, stress in the transitional region of the SiO$_2$ interfacial layer enhances the diffusion of oxygen up from the interface, which makes the $V_t$ roll-off problem more severe. Therefore a stress-
relieved SiO$_2$/Si interface is desirable to minimize V$_t$ roll-off. The SRPO process, deuterium incorporation, and the combination of SRPO with defect passivation with fluorine (see section 3.1.3) are effective approaches to minimize the V$_{fb}$-EOT roll-off problem. Indeed, F+ implanted in the gate stack with 1 nm SiO$_2$ under 2 nm HfSiO$_x$ followed by a 1000°C/10 sec. anneal [62] significantly reduces the roll-off (Fig. 36).

4. Conclusion

The high-k/metal gate stack has been used for high performance and low power semiconductor products replacing the SiO$_2$/polysilicon stack, which has been used for decades. The motivation for this replacement as well as materials screening for the high-k/metal gate stack have been discussed. Two major advantages of implementing HKMG stacks are that they contribute to reducing the gate leakage current and scaling the EOT of advanced CMOS. Progress in high-k/metal gate device threshold voltage tuning, including an evaluation of capping layers, has also been presented. While nMOS V$_t$ is acceptable for various applications of Si CMOS devices, achieving a targeted pMOS V$_t$ is still challenging due to V$_{fb}$-EOT roll-off issues. Approaches to minimize the pMOS V$_{fb}$-EOT roll-off problem have been outlined. Clearly, CMOS scaling will continue to provide opportunities for exciting research in high-k/metal gate modules in the future.

5. References

[3] Donggun Park; Ya-chin King; Qiang Lu; Tsu-Jae King; Chenming Hu; Kalnitsky, A.; Sing-Pin Tay; Chia-Cheng Cheng, “Transistor characteristics with Ta2O5 gate dielectric,” IEEE EDL, P. 441-443, 1998


[23] E. P. Gusev et al., AVS Topical Conf. on Atomic Layer Deposition, May 14, 2001


The Progress and Challenges of Applying High-κ/Metal-Gated Devices to Advanced CMOS Technologies


The evolution of solid-state circuit technology has a long history within a relatively short period of time. This technology has lead to the modern information society that connects us and tools, a large market, and many types of products and applications. The solid-state circuit technology continuously evolves via breakthroughs and improvements every year. This book is devoted to review and present novel approaches for some of the main issues involved in this exciting and vigorous technology. The book is composed of 22 chapters, written by authors coming from 30 different institutions located in 12 different countries throughout the Americas, Asia and Europe. Thus, reflecting the wide international contribution to the book. The broad range of subjects presented in the book offers a general overview of the main issues in modern solid-state circuit technology. Furthermore, the book offers an in depth analysis on specific subjects for specialists. We believe the book is of great scientific and educational value for many readers. I am profoundly indebted to the support provided by all of those involved in the work. First and foremost I would like to acknowledge and thank the authors who worked hard and generously agreed to share their results and knowledge. Second I would like to express my gratitude to the Intech team that invited me to edit the book and give me their full support and a fruitful experience while working together to combine this book.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following: