Advanced high-κ dielectric stacks with polySi and metal gates: Recent progress and current challenges

The paper reviews our recent progress and current challenges in implementing advanced gate stacks composed of high-κ dielectric materials and metal gates in mainstream Si CMOS technology. In particular, we address stacks of doped polySi gate electrodes on ultrathin layers of high-κ dielectrics, dual-workfunction metal-gate technology, and fully silicided gates. Materials and device characterization, processing, and integration issues are discussed.

1. Introduction
The enormous growth of microelectronics over the past four decades and, as a result, the significant progress of information technology in general are based, to a large extent, on a simple gift of nature, the SiO₂/Si system. This is especially true because ultrathin gate dielectrics in MOSFETs remain the key element in conventional silicon-based microelectronic devices. Since the very beginning of the microelectronics era, the SiO₂ gate oxide has played a critical role in device performance and scaling [1–6]. Whereas the thickness of the SiO₂ gate oxide in the first transistors was a few hundred nanometers, the functionality and performance of state-of-the-art devices currently rely on gate oxides that are just a few atomic layers (~1–2 nm) thick. Until very recently, the (evolutionary) scaling of the gate dielectric (and ULSI devices in general) has been accomplished by shrinking physical dimensions. As the physical thickness of SiO₂-based gate oxides approaches ~2 nm, a number of fundamental problems arise. In this ultrathin regime, some key dielectric parameters degrade: gate leakage current, oxide breakdown, boron penetration from the polysilicon gate electrode, and channel mobility [1, 3]. Each of the parameters is vital for device operation. In other words, the conventional device-scaling scenario involving scaling down SiO₂-based dielectrics below 1 nm becomes impractical.

The solution is to replace conventional SiO₂ gate oxides with a material having higher permittivity. High-κ insulators can be grown physically thicker for the same (or thinner) equivalent electrical oxide thickness (EOT), thus offering significant gate leakage reduction, as demonstrated by several research groups [7–10]. Significant progress has been achieved in terms of the screening and selection of high-κ insulators, understanding their material and electrical properties, and their integration into CMOS technology [7–10]. After almost a decade of intense research, the family of hafnium-oxide-based materials, such as HfO₂, HfSi₅O₁₅, HfO₂Nₓ, and HfSiO₃N₂, emerges as a leading candidate to replace SiO₂ gate dielectrics in advanced CMOS applications [11–23]. It has also become evident in the last few years that only replacing the gate insulator, with no concurrent change of electrode material (currently heavily doped polySi), may not be sufficient for device scaling. Polysilicon gate electrodes are known to suffer from a polySi depletion effect (equivalent to a ~0.3–0.4-nm-thick parasitic capacitor), which cannot be ignored for sub-2-nm gate stacks. Therefore, research on dual-workfunction metal-gate electrodes is gaining momentum, since conventional gate stacks are approaching a limit to scaling as a means of improving performance for nano-CMOS (i.e., sub-65-nm) technologies.

It is the purpose of this paper to review our current understanding of advanced metal-gate/high-κ stacks from the perspective of integrating both basic materials and devices. Reliability is also an important factor, especially for long-term device operation. Some reliability aspects of advanced gate stacks are covered in the paper. More
detailed results and focused discussion on this important topic can be found in a recent dedicated review by IBM researchers in [24]. The use of high-κ dielectrics is one of the most critical challenges in CMOS device scaling, and, as such, it is being aggressively tackled worldwide by many researchers and engineers in academic, industrial, and government laboratories. In this review, we focus specifically on the work and progress achieved in the IBM laboratories over the past five years. The paper is organized as follows. In Section 2, we discuss progress and challenges in the integration of high-κ materials with polySi gates. Historically, using polySi gates with high-κ dielectrics was believed to be a “simple” solution to overcome limits to SiO2 scaling in the tunneling regime when gate leakage became too severe. It was the reason why early work in the high-κ area was dominated by devices with polySi electrodes. It was later realized that the polysilicon electrode was not an ideal solution, for reasons of thickness scaling and threshold voltage control; as a result, the focus shifted to metal gates, an even more challenging area. The current status of research and development in this area is reviewed in Section 3. It is demonstrated that metal gates do offer extraordinary scaling potential to an electrical inversion thickness of almost 1 nm. At the same time, dual-workfunction control of n+ and p+ Si band edges remains a challenge. As discussed in Section 4, fully silicided (FUSI) gates combine the integration benefits of polySi devices and metal-like behavior without polySi depletion effects.

2. PolySi/high-κ gate stacks and Hf-based gate dielectrics

PolySi-based devices are usually annealed at high temperatures (>1,000°C) in order to activate dopants in the gate and source/drain regions. The requirements of thermal stability in contact with the polySi gate electrode and negligible metal diffusion into the Si channel have virtually ruled out successful integration of high-κ materials such as ZrO2 [25, 26] and Al2O3 [27] that once were under intense investigation. Even with Hf-based materials, a number of challenges remain, perhaps most significantly regarding the thermal stability of the dielectric, electrical thickness scaling, carrier mobility, p-FET threshold voltage, and long-term stability/reliability under device operation conditions. In the following sections we review how such considerations have recently guided the development of Hf-based gate stack materials for polySi-gated devices.

Thermal stability

In contrast to ZrO2 [25, 26], no detrimental silicide formation occurs with HfO2 in contact with polySi gates. However, polySi/HfO2 gate stacks do undergo substantial changes during thermal processing. Dopant activation requires annealing to temperatures of ~1,000°C or more, much higher than crystallization temperatures of amorphous HfO2. Depending on HfO2 thickness, crystallization into predominantly monoclinic polycrystalline films occurs at 300–500°C [28–31]. Also, the formation of additional interfacial SiO2 is often observed [32], degrading gate stack capacitance. This is discussed in more detail in the next subsection.

In the early stages of work on HfO2, it was believed that grain boundaries in polycrystalline films might constitute electrical leakage paths, giving rise to dramatically increased gate leakage currents. Experimentally, only minor, if any, increase in leakage with polycrystalline HfO2 has been observed [29]. However, amorphous high-κ layers may be preferred for other reasons.

For example, it has been suggested that heterogeneous grain orientations in the dielectric layer may give rise to spatially varying electric fields and thus cause carrier scattering, thereby degrading mobility. Also, with continuing scaling, the gate length will become comparable to the HfO2 grain diameter. According to the International Technology Roadmap for Semiconductors [33], the physical gate length of high-performance devices is projected to reach 18 nm by 2010. This may cause detrimental device-to-device variations in leakage, threshold voltage, etc. Also, integration issues such as line-edge roughness at the bottom of the gate stack during gate stack etch may then arise. We note, however, that so far there is little experimental evidence to support the above concerns.

Finally, grain boundaries in poly- or nano-crystalline material were recently claimed to be responsible for localized unoccupied states below the metal d-state-derived conduction band edge in ZrO2 and other transition-metal (and rare-earth) oxide films [34]. Such defect states have been observed by optical and X-ray absorption spectroscopy as well as by photoconductivity measurements [34]. Indeed, band-edge defect states have recently been shown to occur in HfO2 if and only if the dielectrics exhibit crystallinity as detected by infrared spectroscopy, X-ray diffraction (XRD), and vacuum ultraviolet spectroscopic ellipsometry (VUV–SE) [35]. This is exemplified by the imaginary part ε2 of the dielectric function for HfO2 films grown by atomic layer deposition (ALD), as displayed in Figure 1. When HfO2 thickness (and concomitantly crystallinity) is increased, an absorption feature emerges at ~5.8 eV, i.e., ~0.2–0.3 eV below the bandgap. The same correlation between crystallinity and electronic defects holds also for other HfO2 growth chemistries [35]. These observations are significant from a device perspective, since they may be related to the finding that Frenkel-Poole hopping...
through HfO$_2$ gate dielectrics occurs via trapping sites located a few tenths of an eV below the HfO$_2$ conduction band edge [36]. Ultimately, it is unlikely that such states will be a limiting factor in high-\(\kappa\)-based CMOS technologies, since they line up close to the insulator band edge and are therefore not accessible at the low gate voltages employed in high-performance and low-power technologies. However, at present it is not clear whether additional grain-boundary-induced defect states exist deeper in the bandgap. It therefore appears preferable to employ amorphous high-\(\kappa\) materials.

In order to prevent gate dielectric crystallization and to minimize interfacial SiO$_2$ formation, the thermal stability of the HfO$_2$ dielectric must be increased. This can be achieved by addition of Al [29, 37, 38], Si [31, 35], and/or N [39]. HfAlO gate dielectrics have often been found to reduce carrier mobility, possibly due to fixed charge near the high-\(\kappa\)/channel interface [38]. Therefore, most researchers have recently concentrated on HfSiO and HfSiON.

Substantially increased thermal stability is achieved for example at a comparatively low Si content of Si/(Hf + Si) = 20\%.\(^1\) Even after rapid thermal anneals to 1,000°C for 5 s, such films do not exhibit any infrared phonon modes characteristic of monoclinic HfO$_2$, in contrast to what is observed from as-deposited HfO$_2$ films formed under the same conditions. It is likely that the HfSiO remains mostly amorphous, although partial crystallization into the tetragonal or orthorhombic phase cannot be excluded [31]. However, after longer 900–1,000°C anneals, HfSiO may still crystallize and decompose into HfO$_2$ and SiO$_2$ [28, 31, 39, 40, 41]. Further increased thermal stability can be achieved by additionally introducing N. The tendency to crystallize under extended 1,000°C dopant activation anneals is completely suppressed in HfSiON with a N content of N/(O + N) > 10% [39]. In addition, boron penetration is more effectively prevented by HfSiON than by HfSiO [39, 41].

**Electrical thickness scaling**

In broad terms, the electrical thickness of Hf(Si)O(N)/SiO(N) gate dielectrics is determined by the sum of the electrical thickness of the high-\(\kappa\) layer and the interfacial SiO(N) layer (if present). Therefore, a combination of strategies may be pursued in order to minimize total electrical thickness, each posing its own challenges:

- **Minimize high-\(\kappa\) thickness**, while maintaining a) a closed high-\(\kappa\) layer and b) sufficient Hf content of the gate stack as a whole, ensuring a gate leakage advantage over pure SiON gate dielectrics.

\(^1\)M. M. Frank and L. F. Edge, unpublished work.

- **Minimize the interfacial SiO(N) thickness**, while maintaining a) sufficient/appropriate Si surface functionalization to ensure good (near-homogeneous) high-\(\kappa\) nucleation and hence a closed layer, and b) high carrier mobility.

- **Increase N concentration** in the interfacial SiO(N) and high-\(\kappa\) layers in order to increase the dielectric constant and reduce interfacial layer growth during thermal processing, while maintaining a) low charge trapping and b) high carrier mobility.

In this subsection, we review various surface preparation and process approaches to optimize interfacial layer thickness and high-\(\kappa\) nucleation, and in particular to address the scaling benefit of interfacial nitrogen. The scaling benefit of HfSiO nitridation is discussed below in conjunction with the impact of N on charge trapping and carrier mobility.

Chemically or thermally grown silicon oxide films, preferably with a high density of terminal hydroxyl groups, represent excellent nucleation layers for many ALD- and CVD-based high-\(\kappa\) growth processes [42]. However, their thickness typically ranges from 0.5 to more than 1 nm, contributing significantly to the total electrical thickness. This has motivated the development of alternative surface preparation schemes.

In an attempt to fabricate atomically sharp Si/high-\(\kappa\) interfaces, oxide-free *H-terminated Si(H/Si) substrates* have been utilized. Such H/Si(100) can be prepared quickly and reproducibly by a hydrofluoric acid (HF) wet etch of SiO$_2$/Si, with subsequent water rinse [43]. Such substrates are remarkably resistant to oxidation in laboratory air, and even in O$_2$- or H$_2$O-containing environments at temperatures as high as 300°C. (For a

---

**Figure 1**

Imaginary part \(\varepsilon_2\) of the dielectric function, determined from VUV–SE data, for 5- and 40-nm-thick ALD-grown HfO$_2$ films. The 40-nm data has been offset vertically for clarity.

- **Minimize the interfacial SiO(N) thickness**, while maintaining a) sufficient/appropriate Si surface functionalization to ensure good (near-homogeneous) high-\(\kappa\) nucleation and hence a closed layer, and b) high carrier mobility.

- **Increase N concentration** in the interfacial SiO(N) and high-\(\kappa\) layers in order to increase the dielectric constant and reduce interfacial layer growth during thermal processing, while maintaining a) low charge trapping and b) high carrier mobility.
comprehensive review of H/Si reactivity with respect to O₂, H₂O, NH₃, and high-κ precursors, see [44]. While this low surface reactivity may be advantageous in terms of oxidation resistance, it also causes the poor nucleation characteristics of many ALD-grown high-κ films, resulting in nonlinear growth kinetics and the formation of discontinuous and electrically leaky gate stacks. Prominent examples are the popular HfCl₄/H₂O process for HfO₂ growth [30, 42, 45], as well as other water-based ALD processes employing metal precursors that are designed to react with surface –OH groups, such as ZrCl₄ for ZrO₂ growth [26, 46] and Al(CH₃)₃ for Al₂O₃ growth [44–46]. Nucleation can be enhanced, and more linear growth achieved, if in situ activation of the H/Si surface by a more reactive oxygen precursor such as O₃ is performed. However, this comes at the expense of substantial interfacial SiO₂ formation during growth [47, 48].

A simple way to overcome poor ALD nucleation on H/Si without employing a more reactive O precursor is via initial extended H/Si exposure to Al(CH₃)₃. During exposures ~1,000 times larger than what is commonly employed in ALD, metal–organic functional groups are introduced onto the Si surface [44, 45, 49]. Such groups are reactive toward the water precursor. On Al-organic functionalized Si, improved HfO₂ and Al₂O₃ nucleation is achieved [44, 45, 49]. (Note that, by contrast, large initial H₂O exposures of H/Si leave the H termination nearly unaffected and therefore do not lead to enhanced high-κ growth [44, 45, 49].) A possible shortcoming of Al-organic functionalization is the excessive Al(CH₃)₃ exposure times required with currently available ALD equipment. Also, in view of HfAlO-induced mobility degradation [38], tests must be made to determine whether Al located near a HfO₂/channel interface is acceptable.

Another approach to optimize nucleation and minimize interfacial SiO₂ formation on H/Si is based on ALD growth at reduced temperatures (e.g., 50–100°C). Interfacial SiO₂ formation thus is prevented using both the Al(CH₃)₃/H₂O process for Al₂O₃ growth² and the tetrakis(ethylmethylamino)hafnium/water process for HfO₂ growth² [50].

To passivate the surface and prevent oxidation, hydrogen may be replaced with other atomic species as surface passivant. For example, monolayer chlorine passivation is achieved by a simple Cl₂ gas treatment of H/Si [51], where reaction rates may be enhanced by ultraviolet (UV) light [52]. Preliminary evaluation indicates a minor thickness advantage over SiON interfaces, but nucleation is poor [53].

Like oxide-based substrates, nitride-based interfaces such as high-nitrogen-concentration SiON or pure silicon nitride often are good nucleation layers [30, 54, 55]. In addition, nitridation increases interface permittivity (e.g., k_SiO₂ = 3.9, k_Si3N4 = 7–8) and thermal stability. This scaling benefit can be realized, for example, with interfacial Si(O)N layers fabricated by H/Si anneal in NH₃ at 650°C to form thin Si₃N₄ [55, 56], optionally followed by an oxidizing anneal in NO [57]. PolySi/HfSiO stacks on such high-nitrogen-content films exhibit lower electrical thickness (EOT) than on 1.1-nm low-N-content SiON control substrates, even in cases in which physical thickness is greater [57]. The main concern with high-nitrogen-content interfaces is carrier mobility loss, as discussed below.

To conclude, we note that despite promising results with unconventional Si surface treatments (such as chlorination or metal–organic functionalization), most work in the high-κ field still relies on hydrogen termination and on SiO₂ or low-nitrogen SiON films. Such substrates can be prepared quickly and cost-effectively with conventional manufacturing equipment.

**Carrier mobility**

High-κ materials have often been observed to degrade carrier mobility in the transistor channel. A number of mechanisms have been held responsible, most significantly remote phonon scattering through emission or absorption of low-energy phonons in the high-κ material [58] and remote Coulomb scattering off fixed or trapped charges in the gate dielectric. Over time, reported mobilities with nominally similar gate stacks generally have improved. This suggests that certain defects in the high-κ materials such as electrical trap sites and impurities giving rise to fixed charges can be minimized by process engineering.

However, remote phonon scattering has been central in the debate regarding mobility degradation since, if significant, it could fundamentally limit the performance of HfO₂-based devices. A possible solution is based on the incorporation of Si into the HfO₂, modifying the vibrational properties. Since Si–O bonds are stiffer than Hf–O bonds, soft phonon modes are reduced in intensity. The consequent drop in the remote phonon scattering cross section has been predicted to result in near-complete carrier mobility recovery when ZrO₂ is replaced by ZrSiO₄ [58]. The same physics holds for hafnium silicates, as experimentally proven by Ren and colleagues [59]. The mobility advantage comes at the expense of a reduced dielectric constant (20–25 for HfO₂ compared to 10–15 for HfSiO), a tradeoff that must be taken into account when optimizing overall device performance by tuning the composition of Hf-based dielectrics.

---

Carrier mobility and thermal stability have likely been the main characteristics driving the shift in industry focus from polySi/HfO₂ to polySi/HfSiO(N) gate stacks. Indeed, even Si concentrations in HfSiO as low as Si/(Hf + Si) = 20% have been shown to enable excellent mobilities. For example, we have fabricated polySi/HfSiO/SiON n-FETs with EOT = 1.6 nm that exhibit electron peak mobility identical to that of low-N-content SiON control devices [60]. High-field electron mobility was degraded by only ~10%. Given a leakage reduction factor of >1,000 compared with SiON control devices with the same EOT, such gate stacks are serious contenders for low-power applications.

Whether the observed mobility improvements achieved by the introduction of Si into the HfO₂ are due mainly to the drop in remote phonon scattering is still under debate. Reduced charge trapping is another possible cause, since this would reduce the Coulomb scattering rate. The charge-trapping behavior of HfSiO is indeed better than that of HfO₂ [59].

Nitrogen is often introduced into high-k gate stacks to enhance thermal stability and reduce electrical thickness, as discussed briefly above. However, carrier mobility is usually reduced, e.g., for HfSiO on Si₃N₄ interface layers [56]. This is illustrated in Figure 2, which shows n-FET electron mobility for various HfSiO/SiO(N) gate stacks. Mobility at high field (black symbols) was extracted from full mobility curves (inset) measured using the split C–F technique [61, 62]. With interfacial Si₃N₄ formed by an NH₃ anneal of H/Si at 650°C (N areal density ~2 × 10¹⁵ N/cm²), high-field mobility is degraded by 20–25% compared with low-N-content interfacial SiON layers (~7 × 10¹⁴ N/cm²). Even upon introduction of O into the nitride using NO gas anneals at 700–800°C, mobility recovers only marginally. When interpreting such data, it is noteworthy that high nitrogen concentrations usually reduce mobility even with conventional SiON gate dielectrics [63]. Since nitrogen is known to create fixed charge in SiON [64, 65], it seems natural to hold Coulomb scattering by fixed charges responsible for the mobility loss both in SiON and high-k stacks. However, other physical causes may also underlie the observed N-induced mobility degradation in high-k gate stacks. Three scenarios may explain an observed mobility reduction: a) Slow interface states (areal density Nᵢₛ) or b) fixed charges (areal density Nᵢₓ) cause Coulomb scattering of channel electrons; or c) charge trapping causes Coulomb scattering or induces hysteresis which distorts the inversion charge and mobility measurement.

A combination of electrical measurement techniques aids in assessing which of these mechanisms is dominant in mobility degradation [57]. To address scenario a), Nᵢₛ was measured by amplitude-sweep charge pumping. Independently of O content, all nitride-based interfaces exhibited 3–5 times higher Nᵢₛ (1.3–1.9 × 10¹¹ cm⁻²) than low-nitrogen-content control SiON interfaces. In order to establish whether this Nᵢₛ difference is sufficient to explain the mobility loss, a corrected mobility was calculated that would be measured if Nᵢₛ could be reduced to zero without otherwise modifying the gate stacks.³ After this correction, the mobility trend with N and O content remained virtually unchanged (Figure 2, white symbols), demonstrating that mechanisms other than slow interface states are predominantly responsible for N-induced mobility loss. Scenario b), by contrast, was supported by C–F measurements: threshold voltage Vₜ is ~0.1 V lower with all Si₃N₄-based interfaces than with the control SiON interface. This shift corresponds to an areal density of positive fixed charge of Nᵢₓ ~8 × 10¹¹ cm⁻² (broadly consistent with [56]), independent of O content. Nᵢₓ thus is significantly higher than Nᵢₛ and can quantitatively explain the observed mobility loss with reoxidized nitride interfaces [57]. With pure nitride interfaces, distortion of the inversion charge measurement due to transient charging c) occurs in addition [57]. However, fixed charge likely is the main case of carrier mobility loss with interfacial N [57].

Nitridation of HfSiO layers similarly has often been reported to degrade mobility. However, this is not a universal result. Recent experimental studies indicate that

Electron mobility with 3 nm top-nitrided HfSiON with [N/(N + O)] ~ 21%, compared with low-N-content SiON.

mobility impact is greatest if nitridation conditions allow N to permeate the entire HfSiO film, while near-surface nitridation preserves mobility [39]. This indicates that N in or close to the interfacial layer has by far the greatest impact on mobility. Possible reasons for this are a) the rapidly decaying electrical field strength around a fixed charge, giving rise to simultaneously dropping Coulomb scattering cross sections; and b) a lower fixed charge per N atom in HfSiON than in SiON.

Good electron mobility with HfSiON/SiO2 is demonstrated in Figure 3. Appropriate low-temperature plasma nitridation conditions ensured a high proportion of near-surface N [57]. Under such conditions, gate stacks incorporating such HfSiON showed N-induced $T_{\text{inv}}$ reduction by up to 0.1 nm, confirming the scaling benefit of N. At N concentrations as high as [N/(N + O)] ~ 21%, the N-induced $V_{th}/V_f$ shift to more negative values is smaller than 0.02 V, showing that little positive fixed charge is created far from the gate electrode. Trap density remains low as well. As expected on the basis of our discussion of the mobility degradation mechanisms with interfacial N, mobility is nearly identical to that of a low-N-content SiON control (Figure 3).

Summarizing this section, the replacement of HfO2 by HfSiO has led to mobility improvements, through reduced remote phonon and/or Coulomb scattering. Additional N incorporation helps optimize thermal stability and electrical thickness. However, N near the channel reduces carrier mobility through Coulomb scattering by fixed charges. N incorporation near the top of the HfSiO is therefore the method of choice.

**Threshold voltage**

The threshold voltages of polySi-gated high-κ n-FETs and p-FETs usually deviate from the ideal values achieved with corresponding SiO(N) devices. Using Hf-based high-κ materials, in particular, n-FET $V_f$ is usually found to be more positive by ~0.2 V, while p-FET $V_f$ is more negative by ~0.6 V [66–69]. While threshold voltages can be tuned to their optimum values through device engineering, for example by an appropriate choice of halo implant design or by counterdoping, device performance degrades with excessive tuning. It is likely that n-FET devices can be designed in such a way as to offset the materials-induced shift of ~0.2 V. By contrast, given the ~0.6-V shift for p-FET devices, one cannot rely on implant engineering alone in order to fabricate good-performance Hf-based polySi/high-κ devices. The gate stack itself must be understood and modified. In the following, we first summarize some observations regarding the impact on $V_f$ of processing conditions and materials composition. We then review the current understanding of the underlying physical mechanisms causing the increased p-FET threshold voltage. Finally, we discuss recent attempts to control p-FET $V_f$ and demonstrate that improvement by ~0.3 V can be achieved by appropriate design of the gate stack alone [60], rendering p-FETs with good performance possible.

The fact that laboratories worldwide—using a wide variety of process equipment and chemicals to fabricate gate stacks—report nearly identical p-FET $V_f$ shifts of ~0.6 V from the target value suggests that a fundamental physical or chemical phenomenon is responsible. We have tested whether tuning of processing details, in particular choice of dopant, method of doping (implant vs. in situ doping with CVD precursors), and thermal processing, can help control $V_f$. To this end, the $V_{th}/V_f$ shifts were measured after such critical gate-stack fabrication steps [70]. Measurements even with undoped and unactivated polySi gates were made possible by recording electrical data at elevated device temperatures (up to 200°C) in order to ensure sufficient conductivity. The results indicated that $V_{th}/V_f$ ratios are largely set during polySi deposition and remain virtually unchanged during gate implantation and thermal activation, independent of the p-type dopant (B, Al, Ga). The p-FET $V_f$ shift is thus a fundamental phenomenon that is not easily prevented by employing modified polySi/Hf(Si)O(N) formation conditions. A reaction of Si with the Hf-based material, occurring already during polySi deposition, appears to be the root cause for the poor $V_{th}/V_f$ control.

The introduction of Si or N into the Hf-based layer has a limited impact on $V_{th}/V_{th}$. As expected, when utilizing HfSiO with increasing Si content, $V_{th}$ gradually approaches the value observed with SiO2 (Figure 4, inset) [67, 69–72]. However, in order to bring $V_f$ to within less than 0.3 V from the target value, Hf contents below ~20% are required. At such compositions, the dielectric constant is only marginally higher than for SiON, making
implementation unattractive. As mentioned above, fixed charge from N incorporated into the gate stacks (in particular, into the bottom interface) is another means of controlling \( V_t \). However, only a limited degree of \( V_t \) improvement (by up to \( \sim 0.1 \) V) is achieved in this manner, at the expense of mobility loss [57].

Recent experimental evidence indicates that oxygen plays a prominent role in the p-FET \( V_t \) shift phenomenon. It was demonstrated that oxidation of the polySi/high-\( \kappa \) stack by lateral indiffusion of oxygen can alleviate the p-FET \( V_t \) shift of transistor devices with channel lengths below \( \sim 1 \) \( \mu \)m at the expense of EOT [73]. Also, optical spectroscopy was used to relate trap levels in HfO\(_2\) to oxygen deficiencies [74]. It is likely that these results are related to recent findings for metal/high-\( \kappa \) gate stacks, which are discussed in detail in the next section. There, it is demonstrated that p-FET \( V_t \) can vary by as much as 0.75 V, depending on O\(_2\) partial pressure and temperature during post-deposition gas anneals [75].

Fermi-level pinning has often been invoked as a fundamental mechanism causing the \( V_t \) shift, in analogy with a phenomenon that has long impeded successful fabrication of high-quality gate stacks on compound semiconductors [76, 77]. Fermi-level pinning is caused by a high areal density of interface states whose occupation changes as the gate voltage is swept from conditions of accumulation to inversion. The interface states partially screen the electric field from the gate electrode, preventing it from reaching the channel. The extent of gate-induced tuning of the channel carrier occupancy is thus greatly reduced. In the first detailed discussions of the p-FET \( V_t \) shift with Hf-based high-\( \kappa \) materials [67–69], it was argued that Fermi-level pinning just below the polySi conduction band is caused by Hf–Si bonds at the high-\( \kappa \)/polySi interface. Direct physical evidence for such bonds is scarce, but this picture is broadly consistent with the experimentally observed impact of oxygen deficiencies on \( V_t \).

However, defect levels and fixed charge in the Hf-based gate dielectric itself may similarly cause \( V_t \) shifts. It has been reported, for example, that O vacancy formation in HfO\(_2\) is energetically favorable when the HfO\(_2\) is in contact with a p-doped polySi gate, since such defect states are stabilized by the transfer of two electrons to the gate electrode [74, 78, 79]; this transfer cannot occur in contact with an n-doped polySi gate. Positive fixed charge is thus created inside the HfO\(_2\), shifting the p-FET \( V_t \) to more negative values, which provides an explanation for the experimentally observed \( V_t \) behavior.

More generally, potential physical causes for fixed charge are vacancies or interstitials, foreign atoms such as Si, N, or gate dopants diffused into the high-\( \kappa \) layer. Si and N are not candidate species causing the p-FET \( V_t \) problem, since N-free HfO\(_2\)/SiO\(_2\) stacks suffer from it, and since the intentional introduction of Si partially alleviates the issue (see below). The impact of gate dopants was excluded by the careful experimental studies discussed above [70].

Owing to the accumulating evidence regarding the importance of O, and a better understanding of the electronic structure and formation enthalpy of O vacancies in HfO\(_2\), such vacancies are currently considered to be the most likely origin of the \( V_t \) shifts. However, more physical characterization experiments are required in order to conclusively distinguish such defects from interfacial Hf–Si bonds.

From a technological perspective, it is critical to determine whether the p-FET \( V_t \) can be shifted closer to the target value by choosing appropriate processing conditions. As mentioned above, lateral oxidation of the high-\( \kappa \) layer brings partial relief for short-channel devices [73]. However, the concomitant growth of SiO\(_2\) at the gate electrode interface increases the EOT, O indiffusion and hence \( V_t \) are dependent on channel length, and it is unclear whether the O content of the gate stack can be maintained during the entire device fabrication process. These factors limit the implementation of lateral oxidation.

Motivated by the Hf–Si bond theory, efforts have recently concentrated on thin dielectric cap layers inserted between the Hf-based dielectric and the polySi electrode. However, success with this approach has been mixed, notably weakening the Hf–Si bond theory. For example, Si\(_3\)N\(_4\) [70–72], SiC:H [72], and HfON [80] cap layers lead to only very small \( V_t \) improvement. With SiO\(_2\) cap layers, moderate \( V_t \) improvement (by 0.3 V) has been achieved at a cap thickness of 1 nm [71] (though dissimilar results have been reported [72]). However, SiO\(_2\) capping severely limits thickness scaling and effectively defeats the purpose of introducing high-\( \kappa \) materials.
Recently, aluminum nitride (AlN) was introduced as a novel cap material that reproducibly ensures sufficient $V_t$ improvement at very low cap thickness and high effective permittivity [60, 85]. Hf-based stacks were thus engineered such that the n-FET and p-FET $V_t$ are sufficiently low, with excellent device characteristics. To this end, the AlN cap was deposited onto the HfSiO on both p-FETs and n-FETs, and subsequently etched off the n-FETs. Selective capping of p-FETs only is thus achieved. Separate wafers were employed, but full CMOS integration is possible through a masking/etching scheme.

**Figure 5(a)** shows $C-V$ curves for optimized p-FET and n-FET polySi/(AlN)/HfSiO gate stacks [60]. The physical thickness of the AlN cap is only 0.4 nm; more significantly, because of the high dielectric constant of the AlN, this cap contributes only 0.1 nm to the total EOT, ensuring scalability [85]. The p-FET $V_t$ shift is reduced to only $-0.22 \text{ V}$ to $-0.31 \text{ V}$ compared with a SiON control, depending on the Si/high-$\kappa$ interface layer. For n-FETs, we find $\Delta V_t = 0.21 \text{ V}$, similar to conventional polySi/ Hf(Si)O stacks. Thus, we obtain nearly symmetric $C-V$ characteristics with low $V_t$. These findings are confirmed by $I_d-V_g$ data [**Figure 5(b)**]. A small subthreshold swing of 71 mV/dec indicates that the interface state density is low. This was confirmed by amplitude sweep charge pumping data, which demonstrates that n-FET $D_{it} \approx 10^{10} \text{ eV}^{-1}\text{-cm}^{-2}$ and p-FET $D_{it} \approx 7 \times 10^{10} \text{ eV}^{-1}\text{-cm}^{-2}$ [60]. The p-FET $V_t$ improves slightly with decreasing thickness [60], indicating further scalability. Also, the $V_t$-optimized high-$\kappa$-based FETs show good performance: Mobilities and drain currents for p-FETs and n-FETs range between 90 and 110% of those for a SiON control [60]. A narrow distribution of breakdown voltages indicates the uniform quality of the dielectric. Stress-induced n-FET $V_t$ shifts due to charge trapping are sufficiently low to meet the ten-year device lifetime targets. By combining this capped gate stack with moderate implant engineering for final $V_t$ adjustment, short-channel polySi/(AlN)/HfSiO devices with acceptable performance have been manufactured [85].

In conclusion, O vacancies, or perhaps Hf–Si bonds, in Hf-based polySi/high-$\kappa$ gate stacks are the predominant cause of the observed $V_t$ shifts. It is unclear whether O can be reintroduced into the stack without unacceptable SiO$_2$ growth, and whether such O content can be maintained throughout a full CMOS integration flow. Though unsuccessful in most cases, cap layers deposited onto the Hf-based dielectric have recently shown promise for $V_t$ control. It has been demonstrated that sufficient $V_t$ improvement with a scalable cap layer can be achieved without significantly degrading drive current. In

---

**Figure 5**

(a) Split $C-V$ characteristics for AlN-capped p-FET and AlN-capped/etched n-FET high-$\kappa$ stacks on two interfaces compared with conventional HfSiO and SiON. (b) $I_d-V_g$ characteristics for AlN-capped p-FET (EOT = 1.9 nm) and AlN-capped/etched n-FET (EOT = 2.0 nm) high-$\kappa$ stacks compared with SiON. From [60], with permission; ©2005 IEEE.

Several studies have concentrated on Al$_2$O$_3$ cap layers, often grown by ALD. Reported improvements range from 0.1 to 0.3 V on HfSiO [70, 81, 82] to 0.6–0.7 V on HfSiON [83]. These variations indicate that process control may be an issue. The most likely mechanism of $V_t$ improvement is through negative fixed charge introduced into the Hf-based material by the Al, as is the case for HfAlO gate dielectrics [38]. A possible concern with Al$_2$O$_3$ cap layers is charge trapping under operation conditions, which is a known issue for pure Al$_2$O$_3$ gate dielectrics [14]. In Al$_2$O$_3$/HfSiO stacks, however, the degree of trapping inside the HfSiO appears to be essentially independent of cap thickness [84]. Still, it may be beneficial to reduce cap thickness to a minimum, since with increasing cap thickness the distance of the trap sites from the gate electrode increases and, in turn, the $V_{th}$ shift induced by trapped charges increases.
and mobility requirements have guided the trend of polySi/high-κ devices. In summary, we have reviewed how thermal stability and mobility requirements have guided the trend of polySi/high-κ devices from HfO$_2$ to HfSiO. The incorporation of additional N not only further suppresses high-κ crystallization, but also increases the dielectric constant and aids interfacial layer scaling. However, N close to the channel introduces fixed charge that degrades carrier mobility through Coulomb scattering. Threshold-voltage offset of polySi devices incorporating Hf-based high-κ dielectrics is probably caused by an O deficiency of the gate stack. Scalable AlN capping layers have been developed that enable p-FET $V_t$ control without degrading device performance. Combined with channel engineering by ion implantation, selective p-FET implementation of such AlN/HfSiO gate dielectrics holds promise for successful polySi/high-κ CMOS fabrication.

3. Metal gates

The previous section has shown that while high-κ dielectrics are clearly required to scale beyond the 45-nm node, the integration of Hf-based dielectrics with polySi electrodes suffers from a number of drawbacks, including high p-FET $V_t$ and difficulty in scaling below $T_{\text{inv}}$ of 2 nm. The use of metal gates helps to overcome some of these hurdles. In this section, we summarize the advances and challenges remaining for metal/high-κ stacks. We show that aggressively scaled metal/high-κ stacks ($T_{\text{inv}} = 1.4$ nm) with high electron mobility can be achieved in a conventional self-aligned process by careful process optimization, including the use of non-nitrogen interface layers, high-temperature processing, and appropriate electrode structures to prevent regrowth. However, $V_{th}/V_t$ instability after high-temperature processing remains the biggest challenge to overcome, with oxygen vacancies in the high-κ resulting in large $V_{th}/V_t$ shifts for high-workfunction ($\phi_m$) metal gates.

**Thermal stability**

For compatibility with conventional self-aligned processing, thermally stable metal electrodes were required. This led to our initial evaluation of different metal electrode/dielectric gate stacks by in situ X-ray diffraction (XRD). An electrode was considered unstable if the XRD analysis showed a deviation from the typical linear decrease in diffraction angle ($2\theta$) as a function of temperature [86]. This suggested the reaction and/or formation of a new phase with a different crystal structure. On the basis of this criterion, possible stable electrode choices were narrowed down as shown in Figure 6. Most of the low-$\phi_m$ elemental metal gates ($\phi_m = 4.1$ to 4.3 eV), indicated by light shading, were reactive and did not withstand conventional CMOS annealing temperatures. The exceptions were TaN and TaSiN, which were reported to have low n-FET $\phi_m$ yet remain stable to high temperatures. On the other hand, most of the midgap (including TiN [87], not shown) and high-$\phi_m$ metal gates ($\phi_m = 4.9$ to 5.2 eV), indicated by darker shading, remained stable to high temperatures (800–1,000°C). In summary, while most of the p-FET gate metals and alloys were structurally stable at high temperatures, conventional CMOS processing that requires temperatures greater than 950°C may not be an integration option for most elemental n-FET electrodes.

<table>
<thead>
<tr>
<th>Stability on SiO$_2$ in FG</th>
<th>Stability on Al$_2$O$_3$ in FG</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\leq 700°C$</td>
<td>$\leq 700°C$</td>
</tr>
<tr>
<td>$&gt; 700°C$</td>
<td>$&gt; 700°C$</td>
</tr>
<tr>
<td>IIB IVB VB VIIB VIIB</td>
<td>IIB IVB VB VIIB VIIB</td>
</tr>
<tr>
<td>TaTaN$_3$TaSiN$_3$</td>
<td>TaTaN$_3$TaSiN$_3$</td>
</tr>
<tr>
<td>TaSiN$_3$</td>
<td>TaSiN$_3$</td>
</tr>
<tr>
<td>RuO$_2$</td>
<td>RuO$_2$</td>
</tr>
<tr>
<td>CoSi$_2$</td>
<td>CoSi$_2$</td>
</tr>
<tr>
<td>B C</td>
<td>B C</td>
</tr>
</tbody>
</table>

(a) Periodic table indicating the thermal stability of different electrode materials on SiO$_2$, evaluated using in situ XRD, resistance, and optical scattering analysis techniques. Superscripts following the chemical symbols indicate the type of workfunction: * = n-FET, $\approx$ midgap, and $\approx$ p-FET. The shading indicates whether the thermal stability is less than 700°C (light gray) or greater than 700°C (darker gray). (b) Thermal stability of different electrode materials on Al$_2$O$_3$ (nomenclature similar to part (a)). Adapted from [86], reproduced with permission.
scheme, nitride and oxide are deposited, and this is followed by planarization using chemical–mechanical polishing (CMP). The sacrificial polySi gate and SiON dielectrics are selectively removed, and the new SiON (or high-κ) is grown (or deposited), followed by deposition of the metal gate. After deposition, the highest temperatures to which the gate stacks are exposed are those observed in the back end, which are typically <500°C. Using the replacement-gate integration scheme and CVD W as a metal gate, CMOS transistors down to 0.1 μm were successfully fabricated [88]. It was shown that while the hole mobility of p-FETs remains as good and in some cases better than that of polySi/SiON controls, the electron mobility for W/SiO2, W/SiON [88], and W/HfO2/SiON (Figure 7a) were degraded by more than 20% compared with polySi/SiON gate stacks of similar Tinv. It was also clear that the presence of N in the gate stack further degrades the electron mobility for a low-temperature integration process [88]. Figure 7(b) is a representative TEM image of the center of a 1-μm trench for a W/HfO2/SiON stack. The thickness variation of the HfO2 (intended to be 2.5 nm) clearly demonstrates the conformality issues to be overcome for gate lengths of less than 45 nm. Significant advances have been made with the replacement-gate process [89], in which high electron mobility (250 cm²/V-s at peak and 190 cm²/V-s at 1 MV/cm) at Tinv of 1.5 nm have been observed for a non-nitrogen-based electrode and dielectric process. However, the viability of the replacement gate process for gate lengths of less than 25 nm remains questionable and depends on the development of an extremely conformal dielectric and electrode deposition process.

**Metal/SiON vs. metal/high-κ stacks**

Metal/SiON gate stacks may not be well suited for high-performance logic applications because they do not provide an improvement in leakage and also place severe limitations on deposition processes. Low-damage deposition processes such as ALD or CVD are preferred, since physical vapor deposition (PVD) processes result in sputter damage to the thin oxynitride dielectric layer. Such stringent requirements on deposition processes are not required for the integration of metal/HfO2 gate stacks, since these stacks have a physically thicker high-κ dielectric (compared with SiON) that results in lower gate leakage; more significantly, these stacks are more thermodynamically stable at elevated temperatures than metal/SiO2 stacks [90]. This allows for the possibility of a “gate-first” conventional process integration scheme. Thus, the integration complexity of introducing a metal gate for high-performance CMOS requires that both metal and high-κ be introduced at the same time for an overall benefit to be achieved in scaling and leakage. Unfortunately, aggressively scaled metal/high-κ stacks suffer from electron mobility degradation [87, 91] and Vth/Vi instabilities [75, 92].

**Electron mobility**

In the subsection on thermal stability, it was shown that the low-temperature-processed metal/high-κ devices suffer from degraded electron mobility. The effect of high-temperature processing on the n-FET mobility of W/HfO2/SiO2 stacks was evaluated using a simple non-self-aligned integration flow, with devices processed between 600°C and 1,000°C. It was shown that even with low interface-state densities (Nsi), low-temperature (<600°C) processing resulted in extremely low electron mobilities [93]. Increasing the thermal budget resulted in significantly improved mobilities, but at the expense of Tinv due to interlayer (IL) regrowth. Using a non-self-aligned flow [94], n-FET-like CVD TaSiN/HfO2 gate stacks were also fabricated, and the mobility compared with CVD W/HfO2 after high-temperature processing (Figure 8). As with the replacement-gate results, it is observed that the presence of N at the interface for both gate electrode stacks clearly degrades the electron mobility. This not entirely unexpected, as nitrogen is also the potential cause of the reduction of mobilities for aggressively scaled polySi/SiON devices. For the non-nitrogen ILs, it was not clear whether the observed
improvement in mobility was due only to the thickening of the IL or whether the composition of the IL had also been modified and also played a role. It has been suggested that the Hf intermixes with a non-nitrogen IL (from comparison of high-resolution TEM and electrical measurements) to form a higher-κ Hf-silicate IL [93] that results in higher mobility than HfO₂, since it has a weaker coupling of the SO phonons compared with HfO₂ [95]. However, a number of research groups have chemically analyzed the IL using low-loss electron energy loss spectroscopy [96] and medium-energy ion scattering [97] and show no evidence for Hf-silicate formation upon annealing. Alternatively, it has also been suggested that the IL is Si-rich, resulting in a dielectric constant greater than that of SiO₂ [98]. In summary, the composition of the IL is currently a topic of intense debate in the high-κ community, and its impact on the mobility of high-κ stacks is not well understood.

To understand the effect of processing temperature on electron mobility and to decouple the role of the IL thickening from mobility improvement of metal/HfO₂ stacks, we have used PVD TaSiN, a well-known oxygen diffusion barrier which is known to minimize IL thickening, so that the contribution of the IL to mobility improvement remains constant. For the explicit purpose of dopant activation at low temperature, we used the solid phase epitaxial regrowth (SPER) [99] process, which uses high-energy As implants for S/D amorphization followed by a 600°C anneal, in combination with NiSi S/D and gate contacts to fabricate self-aligned n-FETs at low temperatures. Some wafers were subjected to an additional 800°C, 5 s and 1,000°C, 5 s anneal after SPER and prior to NiSi formation to observe the impact of high-temperature activation. Figure 9(a) shows that substantial improvement in mobility (25%, peak) is observed for both TaSiN/HfO₂/SiON and a control TaSiN/SiON stack only after 1,000°C anneals with little change in $T_{\text{inv}}$. It is clear that the mobility increase is neither related to IL regrowth ($T_{\text{inv}}$ remains about the same; see the figure caption), nor affected by $N_{\text{it}}$ variations, as the mobility curves are corrected for $N_{\text{it}}$ [100]. These results show unequivocally that the high thermal budget modifies the dielectric stack without interfacial regrowth to enhance the mobility. The mobility enhancement can be related to the formation of a relaxed IL/Si interface at $T > 950°C$ [101] or, in addition, especially for the high-κ gate stacks, to structural relaxation and modification of the HfO₂/IL interface.

We have recently obtained high-mobility devices at aggressive $T_{\text{inv}}$, for self-aligned metal-gated high-κ transistors [102] with oxide starting surfaces by capping different thin metal gate stacks such as PVD TiN, ALD TaN, and CVD W with polySi [Figures 9(b), 9(c)]. To prevent reactions between W and polySi at $T > 800°C$, a TiN barrier layer was inserted between the W and polySi layers. PolySi/TiN/HfO₂ gate stacks were shown to have record electron mobilities at a $T_{\text{inv}}$ of 1.4 nm better than previously reported [103, 104]. We believe that by careful process optimization such as the use of non-nitrogen interface layers, high-temperature processing, low $N_{\text{it}}$ ($<3 \times 10^{10} \text{cm}^{-2} \text{eV}^{-1}$), and appropriate electrode and electrode structures to prevent interfacial regrowth, we have largely minimized undesirable sources of Coulomb scattering. This results in high mobility in aggressively scaled metal/high-κ stacks that are competitive or better than aggressive polySi/SiON stacks [Figure 10(a)]. We also show that these high-mobility stacks still maintain more than 4–5 orders of leakage reduction compared with polySi/SiON [Figure 10(b)].

Metal gate screening of the soft optical phonon modes in the high-κ (the primary reason for reduced mobility of polySi/high-κ as proposed by Fischetti et al. [95] and experimentally verified by Ren et al. [59]) has been proposed as a possible reason for improvement in mobility [103]. However, we have recently shown with low-temperature mobility measurements of aggressively scaled metal-gated high-κ stacks⁴ that electron mobility is still limited by HfO₂ SO-phonon scattering.

---

\(T_{\text{inv}}\) scaling

Figures 9(b) and 10(a) show a significant difference in \(T_{\text{inv}}\) for different electrode stacks that are processed under nominally identical process conditions. After a high-temperature process, W-gated devices capped by TiN and polySi are at least 0.5 nm thicker in \(T_{\text{inv}}\) than an equivalent polySi/TiN device. Since the W is completely encapsulated during the S/D activation by TiN/polySi and nitride spacers, the increased \(T_{\text{inv}}\) can only be attributed to residual oxygen present in the W [105] that is released upon annealing as atomic species and oxidizes the Si substrate surface. This kind of regrowth has been observed with other relatively high-workfunction and refractory metals such as Re [75] and suggests that...
possible hurdles may exist for the scaling of high-$\phi_m$ p-FET electrodes.

**$V_t/V_{fb}$ stability—Role of oxygen vacancies**

A key problem that affects metal-gate/high-$\kappa$ stacks is the $V_t/V_{fb}$ stability of metal gates when in contact with Hf-based dielectrics. This remains probably the toughest challenge for the introduction of metal gates and can be summed up as follows: *The $V_t/V_{fb}$ values for metal/high-$\kappa$ stacks predicted using the metal workfunction are accurate for low-temperature-processed devices, but thermal processing induces significant drift, usually toward a midgap effective workfunction (EWF).* An illustration of this effect is shown in Figure 11(a) for PVD TaSiN/HfO2 and CVD Re/HfO2 devices, in which, after high-temperature aneals, the difference in $V_{fb}$ (a measure of the EWF for aggressively scaled devices) is less than 100 mV. The reported $\phi_m$ of these materials is 4.4 eV [106] and 4.9 eV [107]. We reported previously by using the barrier height technique to evaluate $\phi_m$ that some of these $V_{fb}$ shifts that are observed upon annealing are due to fixed charge [108] as interpreted by the difference between the $\phi_m$ obtained from the barrier height technique (which yields values similar to reported $\phi_m$ values) and that extracted from $C-V$. However, it is becoming increasingly apparent industrywide that for high-$\phi_m$ metal gates, the observed EWF on HfO2 can be shifted by more than 500 mV from the expected $\phi_m$ upon exposure to moderately high temperatures and/or reducing ambients. This shift in $V_{fb}$ is qualitatively similar to the high $V_{fb}$ shift observed for p-polySi/HfO2 gate stacks, where the shift was attributed to Fermi-level pinning [68, 69, 109]. Thus, we strongly believe that the $V_{fb}$ modulation is related to the oxygen vacancy concentration [$V_o^\prime$] in the HfO2 near the Re contact.

Recently Shiraiishi et al. [78, 79] have attributed the Fermi-level pinning effect for p-polySi/HfO2 to the generation of an interfacial dipole formed by the evolution of charged oxygen vacancies. By a similar analogy, we believe that the introduction of a high-$\phi_m$ metal gate adjacent to the HfO2 allows for the following reaction: $\text{O}_2 \rightarrow V_o^{++} + 2e^- + \frac{1}{2} \text{O}_2$. Since this reaction is thermally activated, there is no driving force at room temperature for the reaction to proceed (consistent with...
Comparison of $V_f$ instability in polySi, fully silicided (FUSI), and metal gate stacks with the same high-$\kappa$ stressed under identical conditions. Adapted from [115], reproduced with permission; ©2004 IEEE.

The as-deposited e-beam Re/HfO$_2$ measurements. It has been predicted theoretically that the oxygen vacancy defect level in the HfO$_2$ is aligned close to the silicon conduction band [110]. Therefore, at moderately high temperatures, the presence of a high-$\phi_m$ metal with its Fermi level aligned close to the valence band of Si provides the necessary driving force to generate charged oxygen vacancies and lose $2e^-$ to the metal. This results in a dipole layer that changes the effective gate workfunction and the corresponding $V_{th}$ by pulling it toward midgap, as illustrated in Figure 11(d). By introducing oxygen to the system, we can effectively neutralize the oxygen vacancies near the metal/high-$\kappa$ interface, thereby recovering the high $\phi_m$ of the metal gate, as shown in Figure 11(c).

It has also been suggested that these shifts could be attributed to metal-induced gap states (MIGS), an intrinsic effect in which the EWF is modulated by the charge neutrality level and pinning parameter [111], which are well known for HfO$_2$. Recently, Lim et al. [112] have shown quite convincingly that for as-deposited high-$\phi_m$ gate metals, the EWF can be well predicted by the MIGS model; however, upon even moderate annealing the realized EWF can be explained only by the vacancy model. Thus, oxygen movement and its role in modulating oxygen vacancy ($V_o$) formation in the high-$\kappa$ is strongly coupled with the gate electrode and is responsible for the low EWFs that are observed for materials that have high $\phi_m$.

In summary, most of the n-FET metals or alloys are either unstable at high temperatures or at best have EWFs that are more than 200 mV from the Si conduction band edge (for example TaSiN, TaSi$_{2.5}$, or TaC). On the other hand, p-FET metals and alloys, though stable at high temperatures, have unusually high $V_{th}$ shifts that might be related to the oxygen vacancy concentrations in the HfO$_2$ or HfSiO gate dielectric.

**Charge trapping and NBTI**

Unlike polySi/HfO$_2$ stacks which suffer from significant charge-trapping concerns [14, 113, 114], metal/high-$\kappa$ gate stacks have been shown to have very good $V_f$ stability under constant stress conditions; this is illustrated in Figure 12 [115]. Compared with W/HfO$_2$, both FUSI/HfO$_2$ and polySi/HfO$_2$ suffer from significant charge trapping. This degradation is very unlikely to come from the FUSI process, since it is not seen on SiO$_2$ control devices with FUSI gates. These observations combined with the metal gate data strongly indicate that reaction(s) between polySi gates and high-$\kappa$ dielectrics may be responsible for defect creation that leads to enhanced charge trapping, with most of these trapping effects being eliminated by the use of metal gates.

Degradation related to NBTI (negative biased temperature instability) in scaled W/HfO$_2$ replacement-gate p-FETs has also been shown to be comparable to polySi/SiON, suggesting that NBTI is not a problem for aggressive metal-gate/high-$\kappa$ stacks [116].

To conclude this section, substantial mobility improvements in metal-gated high-$\kappa$ systems at an aggressive $T_{inv}$ of 1.4 nm which are as good as or better than those of aggressive polySi/SiON stacks have been achieved. High-temperature processing and nitrogen in the interface layer appear to influence this improvement strongly, though careful process optimization has helped in overcoming mobility as a problem for aggressive stacks. Workfunction stability remains the most significant challenge to overcome, with oxygen vacancies in the high-$\kappa$ resulting in large $V_{t}$ shifts for high-workfunction metal gates. Low-workfunction metal gates are either unstable at high temperatures or are still significantly shifted from the Si conduction band edge. We therefore believe that significant changes to conventional integration schemes would be required in order to obtain high-mobility and band-edge workfunction metal/high-$\kappa$ stacks.

**4. Gate stacks with FUSI metal gates**

As discussed above, using metal gates offers many benefits for CMOS scaling, in particular lower $T_{inv}$ due to eliminated polySi depletion. The process flow described in Section 3 included metallic material deposited directly on gate dielectric regardless of the “gate-first” or “gate-last” integration scheme. An alternative attractive approach to fabricating metallic gates is to convert a conventional polySi gate into a silicide material which, after silicidation transformation, is in direct contact with the dielectric film. Most metal silicide materials are known to have
metal-like low electrical resistivities, typically of the order of 10–100 $\mu\Omega\cdot\text{cm}$ [117–119]. Low resistivity and selectivity to form silicides only in the areas where metal is in direct contact with silicon (the so-called “self-aligned” process) have made them a key contact element of modern ULSI transistors [117]. As a result of many years of focused research and development in this area, silicide materials and processes are fairly well understood.

One should mention that the idea of complete silicidation of a polySi gate was proposed in the late 1970s [117 –120]. At that time, the polysilicon depletion effect was not a big issue, and the focus was more on finding low-resistance contact materials with high reliability. The situation has changed drastically over the past several years with the requirement of reducing electrical thickness of the gate stack in inversion without gate leakage penalty. Several groups have explored full silicidation (FUSI) of conventional polySi gates and observed an encouraging effect of reduced $T_{\text{inv}}$ for the same physical structure and thickness of the dielectric stack [21, 121–144].

Several integration routes to fabricate fully silicided gates have been reported, some involving CMP steps and others not. One popular approach is shown schematically in Figure 13. The integration scheme remains a conventional front-end-of-line (FEOL) process flow including polysilicon gate definition and patterning, ion implantation into extension regions, spacer formation, source/drain ion implantation and silicide contacts, and an oxide passivation layer. After that, FUSI-specific steps include 1) CMP planarization of the passivation overlayer; 2) removal of the cap layer on top of the polySi gate; and 3) metal deposition at the thickness sufficient to fully silicide the polySi gate after moderate annealing, typically at 400–600°C. In this approach, source/drain and gate silicidation are performed separately. To sum up, FUSI gate integration is clearly similar to the conventional CMOS process flow, and therefore offers several advantages over the more complex standard metal gates described above. In fact, short-channel FUSI devices have been demonstrated for 65-nm- and 45-nm-technology nodes.

With respect to silicide materials for FUSI gates, most of the ones explored so far are common silicides that are already in use for source/drain contacts or other microelectronics processes, such as molybdenum silicides [119, 120, 138], tungsten silicides [122], titanium silicides [136], hafnium silicides [134], platinum silicides [131, 133], cobalt silicides [123, 141] and nickel silicides [21, 121, 124–144], germanides, and alloys. Nickel-based silicide materials are emerging as a leading candidate for FUSI gates for several reasons: 1) low resistivity ($\sim 15–25 \mu\Omega\cdot\text{cm}$; 2) low volume expansion (less than 20%); and 3) the fact that this material has already been introduced into Si FEOL processing for sub-90-nm-technology nodes. More significantly, nickel silicide is formed by Ni indiffusion into the polySi gate, therefore allowing complete silicidation without forming voids. As an illustration, a comparison of cobalt silicide and nickel silicide gates is shown in Figure 14. In contrast to nickel silicidation, silicon atoms are the main diffusing species during cobalt monosilicide formation, resulting in
void formation at the gate dielectric interface. Generally speaking, silicidation is a complex multistep process involving diffusion and phase transformation. Depending on the ratio of nickel to silicon, different phases can be formed (e.g., Ni$_3$Si, Ni$_3$Si$_{12}$, Ni$_5$Si$_2$, Ni$_9$Si$_2$, NiSi, NiSi$_2$), with different workfunctions, as is discussed below.

In terms of electrical properties, FUSI gates show a metallic behavior (due to complete silicidation) with no signature of polySi depletion for both high-$\kappa$ and SiO$_2$ gate dielectrics (Figure 15). Accumulation and inversion capacitances are equal, and this is true for both n-FET and p-FET devices. Some slight increase of the capacitance in accumulation is also observed, as expected when polySi gates are replaced with a metal gate. The gain of $T_{\text{inv}}$ due to the FUSI process is approximately 0.3–0.5 nm, especially over the polySi/high-$\kappa$ devices without polySi pre-doping (Figure 15). The combination of polySi-depletion elimination and the high permittivity of the high-$\kappa$ layers results in very significant (six to seven orders of magnitude) gate leakage current reduction, plotted against $T_{\text{inv}}$ (Figure 16). A high-$\kappa$ layer (with polySi gates) contributes to a gate leakage reduction of approximately $10^3$–$10^5$, while FUSI gates offer additional reduction by a factor of $\sim 100$.

As discussed in the two previous sections, threshold voltage control (especially for low-$V_t$ high-performance devices) is a challenge for both metal-gate (band-edge metals) and polySi/high-$\kappa$ devices (the so-called p-FET $V_t$ problem). Achieving band-edge workfunctions for CMOS is one of the key issues with FUSI gates as well. Undoped NiSi gates show a mid-gap workfunction, as evidenced, for example, from $V_t$ shift by $\sim 0.5$ V from n$^+$ Si and p$^+$ Si controls (Figure 15). Several techniques have been proposed to adjust the workfunction of FUSI gates toward band edges: 1) pre-doping of polySi gates with common n$^+$ and p$^+$ dopants before gate silicidation [126, 129–132, 135, 140, 141, 144]; 2) changing the composition of FUSI gates, in particular alloying Ni with other elements (for example, Pt or Ge for p-FET shifts and Al for n-FETs) [130, 131, 133, 137, 140]; 3) using different silicide phases [21, 140, 143, 144];
4) utilizing ultrathin cap materials between the gate dielectric and the FUSI gate, similar to the idea discussed in the polySi/high-κ section [140]; 5) bottom interface engineering [140]; and 6) channel pre-doping.

With the help of polysilicon pre-doping [e.g., As, Sb, P ion implantation (I/I) for n-FETs and Al, B I/I for p-FETs] of FUSI gates on SiO₂-based gate dielectrics, $V_t$ can be adjusted [126] within $150$ meV (for p-FETs) and $300$ meV (for n-FETs) from the mid-gap value of the undoped NiSi (Figure 17). The dopant dose should be carefully optimized because some EOT loss and adhesion problems are observed at high ion implant doses. In other words, there is a tradeoff between the value of the $V_t$ shift and the degree of delamination (for n-type dopants) and also EOT loss. Besides, polySi pre-doping becomes less efficient in the case of FUSI gates on high-κ dielectrics because of the so-called Fermi-level pinning problem discussed in detail in Section 2. For FUSI gates this problem can be mitigated by using 1) metal-rich phases of nickel silicides; 2) platinum silicides or platinum alloys; and/or 3) more stable silicate and nitrided silicate materials. It has been demonstrated that different phases of nickel silicides exhibit workfunctions ranging from $\sim 4.3$ eV (for Ni$_2$Si$_2$) to $\sim 4.7$ eV (for Ni$_3$Si) [21]. This phase-controlled full silicidation offers an extra “knob” to tune the workfunctions of FUSI gates. Another factor in adjusting $V_t$ is to alloy nickel silicides with elements that help to move the workfunction toward band edges. For example, devices with NiPSi FUSI gates show threshold voltages close to a “quarter-gap” p-FET value, whereas alloying with aluminum shifts the workfunction almost to the n⁺ band edge (Figure 18). The mechanism of this $V_t$ modulation is not fully understood at present. It is believed to be possibly due to segregation of the alloying element at the FUSI/dielectric interface.

Device performance improvement is an ultimate goal of device scaling, and innovations in materials and device architecture are enabling it. In terms of performance, long-channel FUSI-gated HfSi$_x$O$_y$ devices show carrier mobilities close to that of the SiO₂ control [131]. This fact combined with reduced $T_{inv}$ (Figures 15 and 16) results in significant drive current improvements [131]. Figure 19 shows (over)drive current in the linear regime as a function of gate leakage. The upper $x$-axis also shows an equivalent gate oxide thickness extracted from gate current density assuming SiO₂ tunneling behavior. At a given gate leakage, the n-FET performance gain is $\sim 25\%$ for NiSi/HfSi$_x$O$_y$ and $\sim 15\%$ for NiSi/SiO₂. Another way
to interpret the data shown in Figure 19 is that, for a given drive current (the equivalent of ~1 nm SiO2 gate dielectric), NiSi/HfSiO shows approximately six orders of magnitude lower gate leakage.

As indicated in the previous sections, charge (electron) trapping is a well-known phenomenon and a serious reliability concern in high-k-based devices. It causes \( V_T \) instabilities and drive current degradation. FUSI-gated devices exhibit charge-trapping behavior similar to that of the polySi/ high-k stacks, as evaluated by means of the constant-stress voltage technique [113]. Specifically, FUSI on HfO2 shows significant \( V_T \) instability, whereas charge trapping in both doped and undoped FUSI on HfSiO\(_x\) is negligible. This important observation was also complemented by charge-pumping measurements.

Finally, we comment on scaling issues of FUSI/high-k gates. “Gate-last” FUSI devices are subjected to processing (starting from polysilicon deposition) and a thermal budget similar to that of polySi/high-k stacks. Hence, one could expect similar issues with regrowth and reactions at high temperatures which should be carefully managed. One conventional way to scale down the electrical equivalent thickness of the stack is to combine an optimized thin SiO2-like interface and a reduced high-k layer thickness. Electrical thicknesses in inversion \( (T_{inv}) \) as thin as 1.6 nm have been achieved for NiSi/HfSiO\(_x\) devices [140].

In summary, the FUSI device is an attractive metal-gate integration option that offers a number of device benefits such as sub-2-nm \( T_{inv} \); performance gain over polySi/SiO2 at a given gate leakage; six to seven orders of gate leakage reduction (at a given \( T_{inv} \)); \( V_T \) control for both n-FETs and p-FETs, and negligible charge trapping.

5. Summary
There is no doubt that enormous progress has been achieved in the area of advanced gate stacks over the past several years. Initial demonstrations of high-k devices in the late 1990s did show significant leakage current reduction due to higher permittivity of the stack. However, these early devices were barely usable. They suffered from significant mobility degradation, threshold voltage instability caused by unacceptable charge trapping, limiting scaling potential below 2 nm \( (T_{inv}) \), reliability concerns, and an unclear integration path. Most of these issues (which seemed fundamental in the early days) have now been solved. High-k/metal-gate devices are much more competitive now for high-performance technologies. They exhibit high mobility at thin \( T_{inv} \) and no significant charge trapping. Controllable and reliable \( V_T \) control still remains as a potential issue, but several options have been identified to solve this problem. Interface optimization is an important task for high-performance (high-mobility) devices.

Acknowledgments
References


35. N. V. Nguyen, M. M. Frank, A. D. Davydov, and D. Chandler-Horowitz, “Sub-Bandgap Defect States in


105. T. Mochizuki, T. Tsujimaru, M. Kashiwagi, and Y. Nishi, “Film Properties of MoSi2 and Their Application to Self-
Evgeni P. Gusev QUALCOMM MEMS Technologies, 2581 Junction Avenue, San Jose, California 95134. Dr. Gusev received his M.S. (applied physics/molecular physics) and Ph.D. (solid-state physics) degrees from the Moscow Engineering Physics Institute (MEPhI) in 1988 and 1991. After graduation, he worked at MEPhI as a Research Associate for two years. In 1993, he joined the Laboratory for Surface Modification at Rutgers University, where he performed research on fundamental aspects of gate dielectrics, first as a Postdoctoral Fellow and then as a Research Assistant Professor. In 1997, he held an appointment as Visiting Professor at the Research Center for Nanodevices and Systems, Hiroshima University, Japan. Dr. Gusev subsequently joined IBM, where he was responsible for several projects related to gate stack processing, characterization, and device integration at both the Semiconductor Research and Development Center (SRDC) in East Fishkill, New York, and the Thomas J. Watson Research Center in Yorktown Heights, New York. In 2005 he joined the QUALCOMM Technology Development Center in San Jose as the Director of the Department of Materials and Device Research and Development. Dr. Gusev has also contributed to the technical R&D community, with nine edited books, more than 140 publications, and 20 issued and filed patents. He is a member of several professional committees, panels, and societies.

Vijay Narayanan IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (vijayna@us.ibm.com). Dr. Narayanan is a Research Staff Member in the Silicon Technology Department at the IBM Thomas J. Watson Research Center. He received his B.Tech. degree in metallurgical engineering from the Indian Institute of Technology, Madras (1995), and his M.S. (1996) and Ph.D. (1999) degrees in materials science and engineering from Carnegie Mellon University, where his dissertation concentrated on understanding the origins of line and planar defects during the epitaxial growth of gallium phosphide on different orientations of Si. In 1999, Dr. Narayanan joined the Department of Chemical and Materials Engineering at Arizona State University as a postdoctoral research associate, with a focus on the initial stages of nucleation and growth of III–V nitrides on sapphire and Si substrates grown by MOCVD. Dr. Narayanan joined IBM in 2001. His current research concerns advanced gate stack technologies including, high-k–metal gate devices for the 45-nm-technology node and beyond. He is an author or co-author of more than 30 peer-reviewed journal and conference papers, and he holds six U.S. patents.

Martin M. Frank IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (mmfrank@us.ibm.com). Dr. Frank is a Research Staff Member in the Silicon Technology Department at the Thomas J. Watson Research Center. He received a Diplom degree in physics from Ruhr-Universität Bochum, Germany, in 1996. He then performed graduate research on oxide-supported metal nanoparticles at Fritz-Haber-Institut der Max-Planck-Gesellschaft in Berlin, Germany, as a scholar of the German National Merit Foundation, and received a Ph.D. degree in physics from Humboldt-Universität zu Berlin in 2000. During a subsequent postdoctoral appointment at Rutgers University, in collaboration with Agere Systems at Lucent Technologies’ Bell Laboratories, he studied dielectric and semiconductor growth on silicon and compound semiconductor surfaces, and metal electrode deposition onto self-assembled monolayers. Dr. Frank joined IBM in 2003. His current research concentrates on high-k gate stacks on silicon and on high-carrier-mobility materials. During an assignment to the Interuniversity MicroElectronics Center (IMEC) in Leuven, Belgium, he also commenced studies of photoresist chemistry. Dr. Frank is an author or co-author of more than 40 papers and one patent. In 2000, he received the Otto Hahn Medal for outstanding scientific achievements.