

Gate contact materials in Si channel devices

Huang-Chun Wen and J.J. Chambers

Research on contact materials in silicon semiconductor devices has recently gained significant momentum due to the increasing performance demands as the complementary metal oxide semiconductor technology advances. Applications include transistor materials such as gate electrodes and contacts to highly doped semiconductors substrates. This review will discuss the key issues in the development of metal gate electrodes with high-κ dielectrics to replace the conventional polycrystalline silicon electrode. Challenges in establishing a work function measurement technique, the role of the metal/high-κ interface in modulating the effective work function, and a review of leading industry solutions will be discussed.

Introduction

For the past few decades, physical scaling of the device and gate dielectric have been the approach to enhance transistor performance.[1] However, traditional silicon dioxide dielectrics have been pushed to the physical limit as the gate oxides are required to scale <1 nm (~3 atomic layers) for transistors having channel lengths below 45 nm. Hafnium-based high-κ dielectrics (HfO_2 and $HfSiO_x$), with a bandgap of ~5.7 eV, are the preferred material system due to their thermal stability with the Si substrate; they also have an appropriate κ value of ~16–25 (Reference 2) to achieve the relatively high capacitance with appropriate physical thickness.[3] The complementary metal oxide semiconductor (CMOS) device is based on the combination of nMOSFET (with electrons as the majority carrier) and pMOSFET (holes). Typical construction of the contact electrode for the two types of devices is differentiated by doping the polysilicon (poly-Si) electrode with *n*- and *p*-type dopants. This modulates the poly-Si effective work function (EWF) at ~4.0 eV and ~5.1 eV (for *n*- and *p*-type dopants, respectively), and enables a low threshold voltage (V_t) for turning on a transistor.

V_t is determined by the flatband voltage (V_{fb}) across the device interface, and V_{fb} is determined by the work function of the metal gate and the silicon substrate.

$$V_t = V_{fb} + 2\varphi_F + V_{ox}, \quad (1)$$

where φ_F is the surface potential in the silicon, and V_{ox} is the potential in the gate dielectric. Furthermore,

$$V_{fb} = \Phi_M - \Phi_S - (Q_f\ \mathrm{EOT})/\varepsilon_{ox}, \quad (2)$$

where Φ_{MS} is the work function difference between gate (Φ_M) and substrate (Φ_S); Q_f is fixed oxide-silicon interface charge density; ε_{ox} is dielectric constant of the insulator; and EOT is equivalent oxide thickness.

Integration of the high-κ dielectrics into the existing process flow with a poly-Si electrode has encountered key roadblocks involving interaction between the poly-Si and the high-κ material. The Fermi-level pinning phenomenon[4,5] suggests the Fermi level of the poly-Si is easily pinned to the energy level of the metal oxide surface states due to charge exchange between the electrode and the dielectric, therefore pinning the EWF of the electrode. The inability to modulate the EWF of the electrode by dopant changes results in an unfavorable increase in the threshold voltage (V_t) for MOSFET devices to turn on (for nMOS, V_{tn} increases ~0.4 eV; for pMOS, V_{tp} increases ~0.7eV).[5] The variables V_{tn} and V_{tp} refer to the threshold voltages for *n*- and *p*-type MOSFET devices, respectively. Another limitation to the use of poly-Si gates is the poly depletion effect while the MOSFET is turned on. This thin dopant depleted layer formed in the poly-Si adds to the EOT (~0.3 nm), which reduces the drain current and cannot be overlooked in technologies targeting <1 nm EOT.[6] These scaling challenges have led to the choice of using metal gate electrodes to replace poly-Si gates.[7]

An early paper from 1997 discussed the need for metal gates in future-generation scaled CMOS devices.[7] However, the choice

Huang-Chun Wen, Texas Instruments; h-wen@ti.com
J.J. Chambers, Texas Instruments; james_j_chambers@hotmail.com
DOI: 10.1557/mrs.2011.8

© 2011 Materials Research Society

https://doi.org/10.1557/mrs.2011.8 Published online by Cambridge University Press

of the appropriate electrode material has not been a trivial task, and the industry has only recently converged on the material and integration strategy.[8–13] Details will be discussed more in the process integration section of this article. The primary reasons for this lack of consensus is that although EWF should be predominantly determined by the bulk/vacuum work function of the materials (materials in columns IIIB, IVB, and VB of the periodic table are suitable for *n*-type EWF; materials in column VIII have *p*-type EWFs),[14] the determination of EWF of a metal gate on high-κ dielectrics exhibits a higher level of complexity, especially the understanding of factors controlling the EWF.

Standardized method for effective work function extraction

The terraced oxide method was proposed to enable the linear extraction of the EWF on high-κ materials from the V_{fb}–EOT relationship[15,16] and to avoid large errors in extracting the EWF from bulk charges in the high-κ material. **Figure 1** shows the terraced oxide methodology, where on a single device wafer, the SiO_2 dielectric is thermally grown (10 nm) and etched into ringed terraces of various thicknesses. Details of this process can be found in References 15 and 16. For EWF extraction on high-κ materials, a thin high-κ film (~2–3 nm) is deposited on the terraced oxide wafers, followed by metal electrode deposition. The terraced oxide method follows a three charge model (considering the fixed interface charge of the Si/SiO_2 interface, the SiO_2/high-κ interface, and minimized high-κ bulk charges) to enable linear V_{FB}–EOT extraction, see reference 16 for more details of the extraction.[16]

With the use of terraced oxide structures, a constant fixed interface charge density Q_f is maintained across the varying oxide thicknesses. The extracted *y*-axis ordinate intercept value is used to extract the metal work function, and the slope is used to extract the fixed charges.

EWF control through metal gate material selection

The effective work function of the metal electrodes can be modulated by changing the material composition (Φ_M term). Pure metals and metal alloys have been demonstrated to exhibit a near linear relationship, with a work function range consistent with bulk work function values.[17,18] However, interface stability of the pure metals with dielectrics is a concern due to additional thermal budgets arising from post gate stack processing. An extreme example of an interfacial reaction dominating the EWF is shown in **Figure 2**. Although Ru on SiO_2 has a high EWF of 4.98 eV, the EWF of Ru on HfO_2 is only 4.53 eV. Careful examination of the Ru/SiO_2 interface shows an interfacial reaction of RuO_x formation, while the interface stability of Ru on Hf-rich high-κ dielectric material is reasonable.[19] Intentional deposition of RuO_2 on a high-κ material shows a high EWF—5.04 eV, confirming that the RuO_x-like interface is the source of the high EWF on SiO_2.

Therefore, the addition of constituents to stabilize the electrode material and provide improved thermal stability with the dielectric is needed. Certain additives to the metal, in particular nitrogen (metal nitrides)[20] and/or Si, Al (metal silicides,[21] ternary

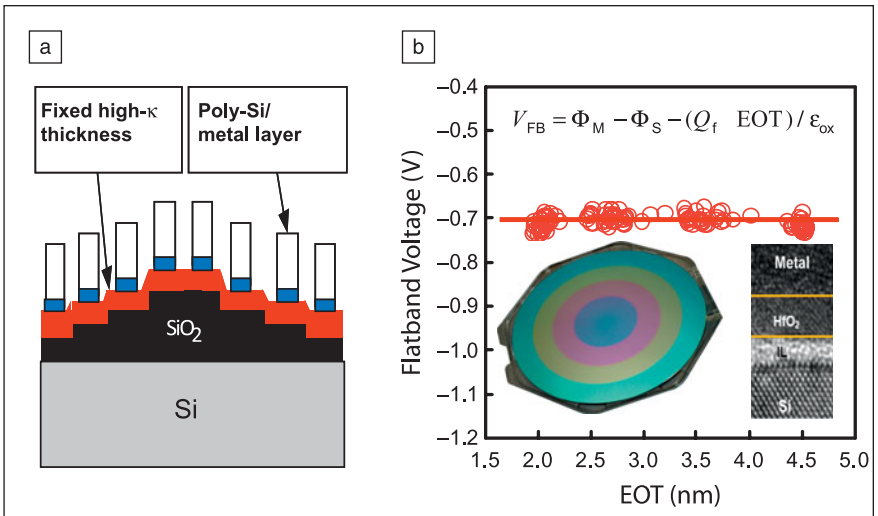


Figure 1. Effective work function (EWF) extraction methodology used to evaluate hundreds of metal gates (see collective data in Figure 2 of the introductory article of this issue). (a) A schematic of the stack showing a fixed high-κ dielectric material on a varying interfacial SiO_2 layer. (b) A flatband voltage versus equivalent oxide thickness curve used to extract the EWF from the *y*-intercept and fixed charges in the stack from the slope. The colored bands on the wafer are regions of different SiO_2 thickness (oxide thicknesses in this image are extremely thick and for visual demonstration only; actual device SiO_2 thicknesses can vary from 10 to 100 Å). The transmission electron microscopy inset shows a schematic cross-section of the typical device.

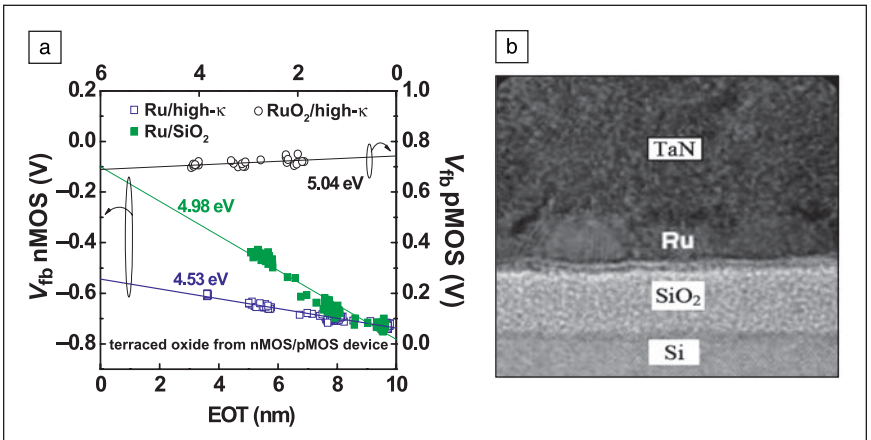


Figure 2. (a) Flatband voltage (V_{fb}) versus equivalent oxide thickness (EOT) from terraced oxide wafers to extract the effective work function of Ru and RuO_2 on SiO_2 and on a high-κ material. (b) High-resolution transmission electron microscopy image of a Ru/SiO_2 interface after a high temperature process showing an interfacial reaction region.

https://doi.org/10.1557/mrs.2011.8 Published online by Cambridge University Press

metal-Si-N,[22] metal-Al-N[23]), have been evaluated and are proposed as possible improvements over pure metals as dual metal electrodes.

An extensive and systematic study of metal EWFs has been reported,[24,25] where the range of EWF tunability for thermally stable metal electrodes is found to be limited to ~4.2–4.9 eV (see Figure 2 in the introductory article of this issue). The interface between the metal gate and high-κ dielectric is found to be critical in determining the EWF, and recent research focus has changed to utilizing interface control for EWF modulation.

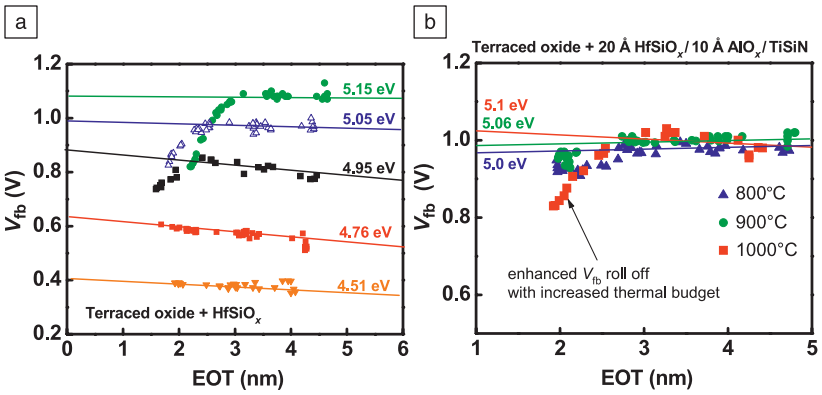


Figure 3. (a) Plot of V_{fb} versus equivalent oxide thickness for high and low effective work function (EWF) metals showing gradual V_{fb} reduction for high EWF metals. (b) V_{fb} roll-off as a function of processing temperature.

EWF control through interface modulation

One such technique is via insertion of a dielectric capping layer between the metal/high-κ dielectric. For engineering toward *p*-type EWF, an Al-based top interface layer (TIL), such as Al_2O_3, or AlN of ~1 nm has been proposed.[26] The EWF values of metal electrodes with Al-TIL increased by ~0.25 eV over their initial values, independent of the metal electrode system. Engineering the interface toward *n*-type EWF is also consistently demonstrated with the use of lanthanide-based dielectrics (La_2O_3, Y_2O_3). The effect is theorized to be due to the formation of a dipole layer at the high-κ/SiO_x interface.[27]

Other techniques include interface modulation by diffusion of additional species (such as Al, La) to the metal/high-κ interface; this utilizes lessons learned from alloyed metals, but can only be implemented in the gate last flow, where the thermal stability requirement is less stringent. Details of this will be discussed in the process integration section of this article. EWF reduction and low V_t NMOS transistors have been reported with the use of a low work function Al cladding layer.[28]

Reports of oxygen control at the metal/high-κ interface have been published, suggesting that an increased oxygen level at the interface can modulate the EWF toward PMOS needs.[29] One hypothesis proposed the cause to be a reduction of oxygen vacancies in the dielectric,[29] but recent reports suggest the increased oxygen level at the interface may be due to the displacement of N in nitride electrodes with O, thus forming an N-induced dipole at the TiN/HfO_2 interface.[30]

V_{fb} or V_t roll-off at low equivalent oxide thickness

One of the most challenging issues in metal gate development is a gradual reduction ("roll-off") in the V_{fb} as the gate stacks are scaled to lower thicknesses for *p*-type electrodes. The onset of "roll-off" occurs at an EOT of ~2 nm (**Figure 3a**). The magnitude of roll-off (ΔV_{fb}) is found to be enhanced with higher EWF (*p*-type) metals and with the increased thermal budget of the gate stacks. One theory of roll-off is its association with the generation of positively charged oxygen vacancies in the interfacial SiO_x layer as oxygen diffuses out into the overlaying high-κ dielectric (**Figure 3b**).[31] The roll-off effect in transistors prevents V_t scaling as EOT declines.

Process integration of metal gate electrodes

The industry is divided between both gate-first [8–10] and gate-last[11–13] high-κ/metal gate integration at the 45 nm and 32 nm device technology nodes. The gate-first integration scheme is characterized by deposition of the high-κ/metal gate stack at the start of the process. This deposition occurs at the same time as the standard polysilicon integration in the currently used process; the high-κ/metal gate stack simply replaces the SiO_2/polysilicon. The high-κ/metal gate stack is then processed through the high temperature (> ~1000°C) of the source and drain activation anneal. On the other hand, the metal electrode in a gate-last integration scheme is deposited at a much later point in the process flow (typically following silicidation of the source and drain), and hence it is only exposed to the low thermal budget (< ~500°C) of the back-end process.

Gate-first flow

Figure 4a depicts cross-sections for typical gate-first integrated devices. The salient features are the dielectric capping layers specifically engineered to create favorable dipoles for setting the EWFs. The metal nitride electrode provides for high carrier densities to eliminate polysilicon depletion effects. The first challenge in a gate-first flow is integration of the dual dielectric capping layers. The first dielectric capping layer (typically Al_2O_3 for the PMOS device) is deposited across the entire wafer. Then the Al_2O_3 is selectively removed from the NMOS regions via a photolithography step and a wet etch step that is selective to the underlying HfO_2 gate dielectric. Next, the NMOS capping layer (typically La_2O_3) is deposited and similarly patterned and wet-etched from the PMOS regions.[32] The metal nitride gate electrode and polysilicon top electrode are deposited, and the entire gate stack must be patterned to define the active gates. The next critical step occurs during the junction activation, which exceeds temperatures of 1000°C. This anneal not only

https://doi.org/10.1557/mrs.2011.8 Published online by Cambridge University Press

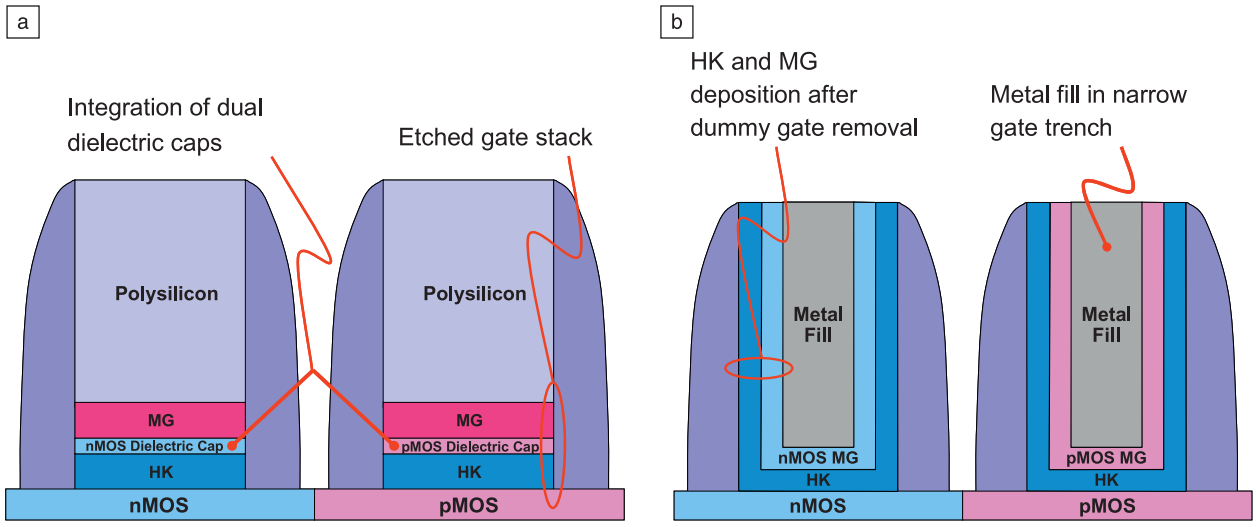


Figure 4. (a) Cartoon depicting a cross-section of typical gate-first integrated nMOS and pMOS devices. Only the gate stack and sidewall are shown for simplicity. The integration of dual dielectric caps and etch of the gate stack are key challenges of the gate-first approach. (b) Cartoon depicting a cross-section of typical gate-last integrated nMOS and pMOS devices. Removal of the dummy gate and metal fill in the narrow gate trench opening are key challenges of the gate-last approach. MG, metal gate; HK, high-κ dielectric.

diffuses and activates dopants in the junctions, but provides the thermal budget to trigger the EWF setting of the dielectric capping layers. However, this step also requires a balance between these beneficial mechanisms and the diffusion of metal from the capping layers to the dielectrics/semiconductor interface, which would cause deleterious V_t shifts and mobility degradation.[32] The rest of the gate-first flow follows the conventional CMOS flow.

Gate-last flow

The main feature of a gate-last flow is deposition of the high-κ/metal gate stack late in the flow benefiting from the low temperature thermal budget of the back-end processes. Some of the key challenges for gate-last integration are illustrated in Figure 4b. The salient feature of the gate-last process is the presence of high-κ dielectric and metal gate electrode films deposited alongside the interior walls of the spacer. This requires the removal of a dummy gate electrode (typically polysilicon) prior to the high-κ/metal gate stack deposition. The removal of the dummy polysilicon is challenging, requiring chemical-mechanical polishing to expose the dummy gates[11] and selective removal (a wet etch process utilizing NH_4OH) of the silicon dioxide underneath the dummy gate. An atomic layer deposition process for high-κ dielectric and the nMOS and pMOS metal electrodes deposition is needed to ensure highly conformal deposition at the bottom of the gate stack. The critical step at this point in the flow is the selective patterning of the opposite electrode material from its complementary gate, which requires challenging photolithography and etching of the metal gate from the narrow gate trench. The gate stack formation is completed with the deposition of the metal fill material (typically Al). A final planarization of the surface using chemical-mechanical polishing completes the device formation.[11]

When contrasting the advantages and challenges of gate-first versus gate-last processes, the control of the EWF is the deciding factor. Since the gate-last flow only exposes the gate stack to the low thermal budget of the back end, the EWF can be set and controlled to the desired band-edge target. This is especially important for the pMOS device, which has typically been susceptible to work function roll-off at high thermal budget.[25] EOT control is another important aspect, since the use of dielectric capping layers in the gate-first flow results in a larger electrical thickness (~2 Å) on pMOS than nMOS.[10] This is not the case for the gate-last flow, where the same EOT can be achieved on both nMOS and pMOS structures. The gate-last scheme is becoming a favorite among many device manufacturers.

Summary

The optimal metal gate material should achieve band-edge work function for low MOSFET threshold voltage control, while maintaining thermal stability on the high-κ dielectric. The key to this finding is that although material selection is important, the critical factor to attaining band-edge work function is in the structure and composition control at the high-κ/metal gate interface. Convergence in the industry in recent years has shown that interface engineering is the approach for "gate-first" process integration strategies, while a new low thermal budget "gate-last" integration process has made high-κ/metal gate technology possible.

References

1. G.E. Moore, "Progress in digital integrated electronics," presented at Electron Devices Meeting, International, 1975.
2. B.H. Lee et al., "Ultrathin hafnium oxide with low leakage and excellent reliability for alternative gate dielectric application," presented at Electron Devices Meeting, 1999. *IEDM Technical Digest, International*, 1999.
3. G.D. Wilk, R.M. Wallace, J.M. Anthony, *Journal of Applied Physics* **89**, 5243 (2001).

https://doi.org/10.1557/mrs.2011.8 Published online by Cambridge University Press

4. C. Hobbs, L. Fonseca, V. Dhandapani, S. Samavedam, B. Taylor, G. Grant, L. Dip, D. Triyoso, *VLSI Technology Digest*, 9–10 (2003).
5. Y.C. Yeo, P. Ranade, T.J. King, C.M. Hu, *Journal of Applied Physics* **92**, 7266 (2002).
6. C.Y. Wong et al., "Doping of n+ and p+ polysilicon in a dual-gate CMOS process," presented at Electron Devices Meeting, 2003.
7. J.C. Hu, H. Yang, R. Kraft, A.L.P. Rotondaro, S. Hattangady, W.W. Lee, R.A. Chapman, C.-P. Chao, A. Chatterjee, M. Hanratty, M. Rodder, I.-C. Chen, "Feasibility of using W/TiN as metal gate for conventional 0.13 μm CMOS technology and beyond," Electron Devices Meeting, 1997. *Technical Digest, International*, 825 (1997).
8. M. Chudzik, B. Doris, R. Mo, J. Sleight, E. Cartier, "High-performance high-k/metal gates for 45 nm CMOS and beyond with gate-first processing," VLSI, 194, 2007.
9. X. Chen, et al., "A cost effective 32 nm high-K/ metal gate CMOS technology for low power applications with single-metal/gate-first process," VLSI, 88, 2008.
10. F. Arnaud, et al., "32 nm general purpose bulk CMOS technology for high performance applications at low voltage," IEDM, 633 (2008).
11. K. Mistry, et al., "A 45 nm logic technology with high-*k*+metal gate transistors, strained silicon, 9 Cu interconnect layers, 193 nm dry patterning, and 100% Pb-free packaging," IEDM, 247, 2007.
12. P. Packan, et al., "High performance 32 nm logic technology featuring 2nd generation high-k + metal gate transistors," IEDM, 659, 2009.
13. C.H. Jan, et al., "A 32 nm SoC Platform Technology With 2nd Generation High-k/Metal Gate Transistorspp," IEDM, 647, 2009.
14. H. Michaelson, *Journal of Applied Physics* **48**, 4729 (1977).
15. H.C. Wen et al. *IEEE Electron Device Letters*, **27**, 598–601 (2006).
16. G. Brown et al., "An Improved Methodology for Gate Electrode Work Function Extraction in SiO2 and High-k Gate Stack Systems Using Terraced Oxide Structures," *IEEE SISC conference*, 2004.
17. H.C. Wen et al., "Work function engineering of RuHf alloys as gate electrodes for future generation dual metal CMOS," presented at VLSI Technology, 2005.
18. H.C. Wen, PhD dissertation, The University of Texas at Austin, pp. 61, 130 (2006).
19. H.C. Wen, P. Lysaght, H.N. Alshareef, C. Huffman, H.R. Harris, K. Choi, Y. Senzaki, H. Luan, P. Majhi, B.H. Lee, M.J. Campin, B. Foran, G.D. Lian, D.-L. Kwong, *Journal of Applied Physics* **98**, 043520 (2005).
20. H.-C. Wen, K. Choi, P. Lysaght, P. Majhi, H. Alshareef, C. Huffman, R. Harris, H. Luan, B.H. Lee, N. Yamada, presented at VLSI Technology, 2005. (VLSI-TSA-Tech). 2005 IEEE VLSI-TSA International Symposium on pp. 107–108, 2005.
21. J.H. Sim, et al., *Electron Device Letters, IEEE*, **24**, 631 (2003).
22. H.C. Wen, et al., *VLSI Technology, 2005. Digest of Technical Papers. 2005 Symposium*, 46 (2005).
23. H.C. Wen. et al., *VLSI Technology, 2007. Digest of Technical Papers. 2007 Symposium*, 160 (2007).
24. H.C. Wen, *Microelectronic Engineering* **85**, 2 (2008).
25. H.N. Alshareef et al., *Future Fab* **19**, 91 (2006).
26. H.N. Alshareef, H.F. Luan, K. Choi, H.R. Harris, H.C. Wen, M.A. Quevedo-Lopez, P. Majhi, B.H. Lee, *Applied Physics Letters* **88**, 112114 (2006).
27. P.D. Kirsch et al., *Applied Physics Letters* **92**, 092901 (2008).
28. C. Hinkle et al., "Dipole controlled metal gate with hybrid low resistivity cladding for gate-last CMOS with low Vt," VLSI Symp 2010 17.3.
29. S.C. Song et al., "Mechanism of V_{FB} roll-off with high work function metal gate and low temperature oxygen incorporation to achieve PMOS band edge work function," *Electron Devices Meeting, 2007. IEDM 2007. IEEE International*.
30. C. Hinkle, R.V. Galatage, R.A. Chapman, E.M. Vogel, H.N. Alshareef, C. Freeman, E. Wimmer, H. Niimi, A. Li-Fatou, J.B. Shaw, J.J. Chambers, *Applied Physics Letters* **96**, 103502 (2010).
31. G. Bersuker, C.S. Park, H.C. Wen, K. Choi, O Sharia, O., A. Demkov, "Origin of the flat-band voltage (V_{FB}) roll-off phenomenon in metal/high-k gate stacks" *ESSDERC* 2008.
32. K. Tatsumura et al., "Intrinsic Correlation between Mobility Reduction and Vt shift due to Interface Dipole Modulation in HfSiON/SiO_2 stack by La or Al addition," in *IEDM Technical Digest*, IEDM, p. 25 (2008). □

Handbook of Modern Ion Beam Materials Analysis
Second Edition



EDITORS
Yongqiang Wang
and Michael Nastasi

NOW AVAILABLE
The most comprehensive database on ion beam analysis ever published—revised and updated from the popular handbook released in 1995!

MRS Visit www.mrs.org/IBH2 for details.

JANIS

UHV Micromanipulated Probe Station



- UVH environment for electro-optical probing of wafers, chips, nanoscale devices
- Vacuum level: ~5 x 10^{-8} Torr at 290 K and ~5 x 10^{-9} Torr at 5 K
- Ultra-low vibrations (<2 nm/min position drift) and wide temperature range (4 - 450 K)
- Rapid sample change via load lock with proprietary demountable cryogenic sample holder
- Two to six probes, DC to 67 GHz, plus fiber optics
- Complete package includes temperature controller, zoom monoscope, integrated turbopump, vibration isolation stand

Janis Research Company
2 Jewel Drive Wilmington, MA 01887 USA
TEL +1 978 657-8750 FAX +1 978 658-0349 sales@janis.com
Visit our website at www.janis.com

www.mrs.org

https://doi.org/10.1557/mrs.2011.8 Published online by Cambridge University Press