Dual Work Function Metal Gate Technology for Future CMOS Devices

Pushkar Ranade, Qiang Lu, Igor Polishchuk, Hideki Takeuchi, Chenming Hu and Tsu-Jae King

Departments of Materials Science and Engineering, Electrical Engineering and Computer Sciences,
University of California, Berkeley, CA 94720

ABSTRACT

Polycrystalline silicon (poly-Si) has been used as the gate material for MOSFETs for several decades. This is because it is highly compatible with CMOS processing, and its work function can be selectively modified by ion implantation of the appropriate dopants. The gate-depletion effect, which increases the equivalent SiO_2 thickness (EOT) of the gate dielectric by several Angstroms and thereby degrades transistor performance, becomes a significant issue for poly-Si gate technology as the EOT is scaled to 1 nm and below in sub-50 nm devices, however. Because of this, alternative gate-electrode materials which can either reduce or eliminate the gate depletion effect have been investigated by many researchers. While refractory metals and their metallic derivatives are attractive candidates, there are numerous process integration challenges that must be overcome before a viable metal-gate CMOS technology can be developed. This paper highlights some of these challenges and reviews three different process integration schemes.

INTRODUCTION

As the gate lengths of MOSFETs are aggressively scaled down to below 50 nm, novel gate-stack materials will be needed to ensure that transistor performance specifications are met. According to the *International Technology Roadmap for Semiconductors* (*ITRS*) [1], CMOS devices beyond the 70 nm technology node will require equivalent oxide thickness (EOT) below 1 nm. Since direct-tunneling current becomes unacceptably high for SiO_2 when it is scaled to less than 1 nm thickness, high-permittivity (high-k) gate-dielectric materials (which can provide much lower gate-leakage current than SiO_2 for the same EOT) have been actively investigated. Devices with less than 1 nm EOT high-k dielectrics have already been demonstrated by a number of research groups. Poly-Si gate depletion can contribute several Angstroms to the EOT, resulting in reduced transistor drive current as well as degraded short-channel effects. The gate-depletion effect is a serious issue, particularly for devices with sub-1 nm EOT. The use of metal gate materials eliminates this problem, and is of most benefit for devices with very thin EOT, as recently demonstrated with SiO_2 and silicon-nitride gate dielectrics [2, 3].

In addition to the gate depletion effect, incompatibility with high-k gate dielectrics is also an issue for poly-Si gate technology. There have been successful demonstrations of MOS capacitors and transistors with sub-1 nm EOT using high-k gate dielectrics viz. oxides of Hf, Zr, Al and La [4-6]. Many of these materials are not thermally stable in contact with poly-Si above certain temperatures, and therefore cannot be used in a conventional dual poly-Si gate CMOS process, in which the highest annealing temperature can exceed 1000°C. Therefore, alternative gate materials that are thermally stable with various advanced gate-dielectric materials will be needed.

For thin gate dielectrics, boron penetration into the channel is a serious problem for poly-Si gate technology. It occurs even with high-k dielectrics [7], and can be eliminated with the use of metallic gate electrodes.

Metal gate materials appear to be essential for future CMOS devices, because of these issues for poly-Si gate technology. However, there are significant process integration challenges that must be overcome before a metal gate technology can be considered to be viable for integrated-circuit manufacturing. These are discussed in the following section.

MATERIALS AND PROCESS INTEGRATION ISSUES

Candidate gate materials must satisfy several criteria in order to be viable for use in a Si-based CMOS fabrication process. For sub-70 nm technology nodes, the *ITRS* specifies a gate sheet resistance of ≤5 ohms/square. Thus these materials must be highly conductive and must have very high melting points in order to withstand the high thermal budgets commonly used in CMOS processing. It is also important that these materials lend themselves easily to conventional thin-film deposition (physical or chemical vapor deposition) and reactive ion etching (RIE) techniques. These requirements will ensure that metal gate CMOS devices can be fabricated using conventional tools. The metal gate materials also need to have thermal expansion coefficients that closely match those of the single crystalline Si substrate in order to ensure that no significant thermal stresses are introduced in the film during rapid temperatures changes (as used for dopant activation). The general requirements described above already limit the candidate materials to some of the high melting point refractory metals, *e.g.* W, Ti, Ta, Mo, Nb and their binary or ternary metallic derivatives, *e.g.* WN, TiN, TaN, MoN, MoO_2. The most significant constraint in the choice of gate material, however, relates to the need to precisely engineer the transistor threshold voltages. To obtain low and symmetric n-MOSFET and p-MOSFET threshold voltages (V_T) while suppressing

short-channel effects, it is essential to have the gate work function between ±0.2eV of E_C and E_V for bulk-Si n- and p-MOSFETs, respectively [8]. On the other hand, advanced transistor structures such as the ultra-thin body silicon-on-insulator (SOI) MOSFET [9] and the double-gate MOSFET [10] that use undoped Si channels require the gate work functions to be between ±0.2eVof E_i, the intrinsic Si Fermi level [11]. There is thus the need to identify metallic materials that have work functions close to the above values and to develop process integration schemes that will provide two different metal gate work functions for NMOS and PMOS devices integrated on a single Si substrate. A further challenge is imposed by the dependence of the metal work function on the properties of the underlying dielectric film. In general, metal work functions at dielectric interfaces differ from their values in vacuum [12]. This indicates that the search for metal gate materials must be conducted in tandem with the search for alternative gate dielectrics. It should also be noted that metals with complementary work functions also display inherent differences in some physical properties, viz. reactivity. Low work function metals are typically easily oxidized while high work function metals are inherently inert and difficult to etch. This imposes additional constraints on the choice of metal gate material and process integration.

In general, there are two major approaches to integrating novel gate-stack materials into a CMOS process. One is the "gate-first" approach, where the gate stack is formed before the source and drain, as in a conventional CMOS process. The other is the "gate-last" approach, where the gate stack is formed after source and drain formation. An example of the latter is a replacement gate process with chemical-mechanical polishing (CMP) of the metal gate [13, 14]. This process is difficult and relatively complex, and has yet to be applied to a full CMOS process. The gate-first approach is generally preferred for its simplicity and compatibility with existing CMOS process flows. Therefore, the gate-first approach is emphasized in this review. Two gate-first processes using dual metal gates, as well as a technique of achieving tunable gate work function using a single metal gate material, are discussed in the following sections.

DUAL-METAL GATE CMOS

This approach derives its name from the fact that two metals are used independently on a single substrate to form the complementary gate electrodes. This process involves the deposition of a first metal layer over the Si substrate, selective removal of this first layer from either the n- or p-well regions (Fig. 1), and subsequent deposition of a second metal layer. This approach ensures that the n- and p-MOSFET threshold voltages are determined by two different metal gate materials.

In the first demonstration of this approach, Ti and Mo (each capped with a barrier layer of TiN) were selected as the NMOS and PMOS gate-electrode materials, respectively [15]. Ti is known to be a low work function metal, while Mo displays a high work function for certain crystallographic orientations [16]. It should be noted that the gate dielectric used was Si_3N_4, because Ti is known to be thermally unstable on SiO_2. N^+-doped poly-Si was deposited over the metal layers to provide sufficient thickness and mechanical stability to the gates. After a multi-step reactive ion etch (RIE) to etch through the composite gate stack, the device cross-sectional views were as shown in Fig. 2. Device fabrication was completed using a

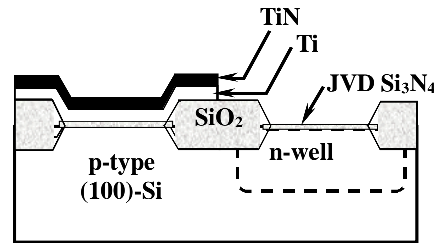

Fig. 1 Cross-sectional view of the dual metal gate CMOS devices after selective removal of Ti from the PMOS regions, before PMOS gate electrode (Mo) deposition.

conventional CMOS process sequence. Normal n- and p-MOSFET device characteristics were adhieved, as shown in Fig. 3. The gate work functions were determined by matching measured capacitance *vs.* voltage (*C-V*) characteristics to those obtained by simulation accounting for quantum-mechanical effects, and were 4.56 eV for NMOS and 4.72 eV for PMOS. These values are higher (lower) than the expected Ti (Mo) work functions in vacuum by ~0.2eV. It is hypothesized that high-temperature annealing might have caused intermixing between the Ti and TiN barrier layer to give a higher work function (intermediate between Ti and TiN). As mentioned earlier, metal work functions at dielectric interfaces are also affected by the

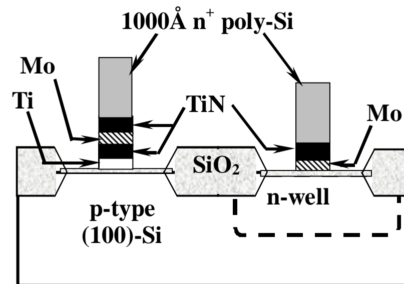

Fig. 2 Cross-sectional view of the dual metal gate CMOS devices after gate stack etch. A conventional process sequence was subsequently used to finish the CMOS devices.

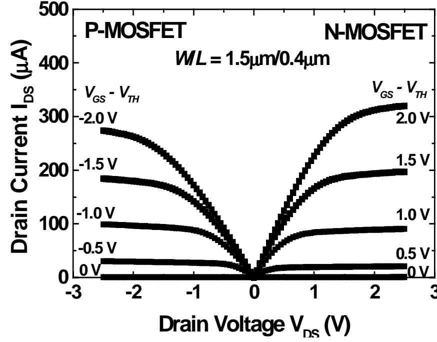

Fig. 3 Normal transistor characteristics were obtained on both p- and n-MOSFETs fabricated using dual metal gates.

nature of the gate dielectric. Theoretical prediction of the work function of (110)-oriented Mo at the Si_3N_4 interface indicates a value close to 4.7eV, consistent with the value obtained in this work [12].

During the selective removal of the metal from the PMOS regions, the thin dielectric film is exposed to the metal etchant. Because Si_3N_4 is slightly etched by the Ti etchant, the PMOS EOT ended up to be ~9Å thinner than that of the NMOS. Such a difference in EOT is a potential reliability concern. This problem could potentially be alleviated through the use of a more selective metal etch process, and could be reduced for alternative gate dielectrics and/or metals.

METAL-INTERDIFFUSION-GATE (MIG)

This approach is so named to indicate the fact that unique segregation and interdiffusion phenomena between thin metal film stacks are exploited to engineer the gate work function. The premise of this approach is that if a bi-layer (or tri-layer) metal stack is used as the gate electrode and the individual layers are thin enough, a suitable thermal budget can be identified that will cause the metals to interdiffuse. Depending on the particular material system used, the metal interdiffusion can result either in a composite alloy film with a unique work function or segregation of the top metal film to the bottom (gate-dielectric) interface. In the latter case, the work function of the gate stack will be that of the metal which segregates to the dielectric interface.

This approach was first demonstrated using Ni and Ti as the two metal films [17]. Ti was selected to be the NMOS gate material because of its low work function, while Ni was selected to be the PMOS gate material. The fabrication process used is similar to the one

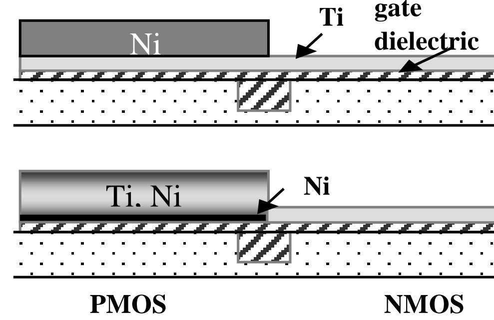

Fig. 4 Cross-sectional view of MIGFET devices (top) before and (bottom) after the metal interdiffusion anneal.

used in the dual-metal gate approach described previously. However, in the MIG process, the bi-layer stack is deposited over the entire wafer. Then, only the top metal (Ni) is removed from the PMOS (n-well) regions, while the stack remains

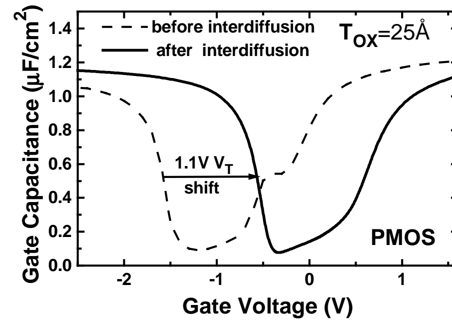

Fig. 5 *C-V* curves of a Ni/Ti stacked gate p-channel MIGFET. The interdiffusion of Ni to the gate-dielectric interface results in a large shift in the flat-band voltage.

undisturbed in the NMOS (p-well) regions. Thus, the gate dielectric is always protected with a metal layer and is never exposed to etchants. After the selective removal of Ni from the NMOS regions, the n-MOSFET gate work function is determined by the only remaining metal (Ti). A low-temperature interdiffusion anneal (400°C, 10 min) then causes the Ni and Ti bi-layers in the PMOS regions to intermix, resulting in Ni segregation to the upper and lower (gate-dielectric) interfaces. The p-MOSFET gate work function is therefore determined by the segregated Ni layer. The key process steps for forming the dual metal gates are shown in Fig. 4. The large shift (1.1 eV) between the p-MOSFET *C-V* curves (Fig. 5) before and after interdiffusion anneal is indicative of Ni segregation to the gate-dielectric interface. This approach thus overcomes some of the limitations of the previous approach while giving the desired n- and p-MOSFET gate work functions. It should be noted that Ti is not an appropriate choice for metal gate material if SiO_2 is used as the gate dielectric in a conventional (gate-first) process, because of thermal instability.

SINGLE-METAL TUNABLE-WORK FUNCTION

Either of the aforementioned approaches can be used to provide dual metal gate work functions on a single Si wafer. However, a technique which requires the deposition of only one metal with a tunable work function would be much more desirable, because it would simplify process integration as well as provide a means of fine-tuning V_T to optimize the trade-off between transistor drive current and leakage. Such a tunable work function approach has been demonstrated using Mo as the gate metal [18, 19]. Mo was selected for its excellent compatibility with Si processing and attractive properties including high melting point, low resistivity and a coefficient of thermal expansion closely matching that of Si. Mo displays a significant anisotropy in work function, with the densely

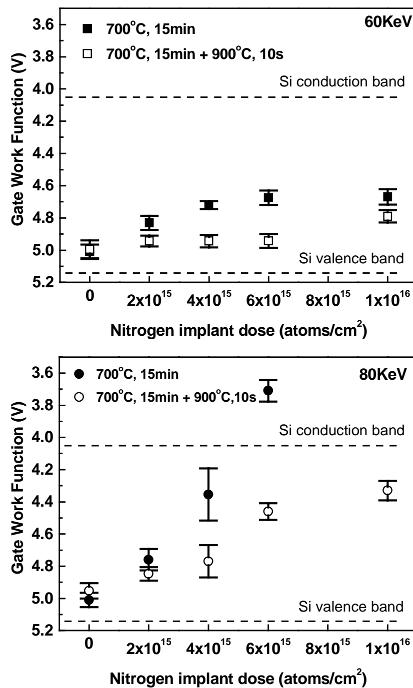

Fig. 6 Impact of nitrogen implant dose and energy on the work function of Mo gate films on SiO_2 gate dielectric [18].

packed (110) planes displaying a high work function (~5eV) and open (112) planes displaying a low work function (~4.3eV) [20]. As mentioned earlier, high work function metals are typically difficult to pattern using reactive ion etch (RIE) processes. Mo thin films, when deposited under certain conditions, can be made to grow with a columnar morphology with (110) planes parallel to the substrate. This leads to a high work function at the interface while allowing for efficient RIE based on Cl or F chemistries. The work function of Mo is suitable for bulk-Si p-MOSFETs, and can be controllably lowered by ion implantation [18]. As shown in Fig. 6, N implantation into the Mo films followed by thermal annealing can lower the Mo work function to ~4.3eV. This change is believed to be primarily the result of crystallographic and chemical changes induced by substantial N segregation at the dielectric interface upon annealing. The high dose N implant initially amorphizes the film, breaking up the grain distribution and lowering the interfacial work function. During subsequent annealing, the film begins to recrystallize, leading to an increase in the work function. The regrowth of the film morphology is however accompanied by the formation of MoN at the interface, which restricts the work function increase. High-energy synchrotron XRD analysis indicates the formation of (211) and (103) oriented MoN [20]. The work function of this particular orientation of MoN is undocumented, however, it is quite likely that these open configurations have substantially lower work functions. N implantation is thus a promising technique for engineering the work function of Mo over a wide range, making Mo suitable as a metal gate material for bulk-Si and SOI CMOS devices.

SUMMARY

Metal gate technology will be necessary to overcome the limitations of poly-Si gate technology, in order for highly scaled CMOS devices to meet performance specifications in the future. Various process integration issues must be considered in developing a viable metal gate technology. Several approaches to achieving dual work function metal gate CMOS have been demonstrated and are reviewed in this paper.

REFERENCES

1. *The International Technology Roadmap for Semiconductors*, Semiconductor Industry Association, 2000
2. R. Chau *et al.*, *IEDM Tech. Dig.*, p.45 2000.
3. B. Yu *et al.*, *Symp. VLSI Tech.*, p. 9, 2001.
4. C.H. Lee, *et al.*, *IEDM Tech. Dig.*, p.27, 2000.
5. B.H. Lee, *et al.*, *IEDM Tech. Dig.*, p.39, 2000.
6. R. Choi, *et al.*, *Symp. VLSI Tech.*, p.15, 2001.
7. K. Onishi *et al.*, *Symp. VLSI Tech.*, p.131, 2001.
8. I. De et al, *Solid State Electronics*, **44**, 1077, 2000.
9. Y.K. Choi *et al.*. *IEDM Tech. Dig.*, p.919, 1999.
10. Y.K. Choi *et al.*, *IEDM Tech. Dig.*, p.421, 2001.
11. L. Chang *et al.*, *IEDM Tech. Dig.*, p.719, 2000.
12. Y. C. Yeo *et al.*, *Symp. VLSI Tech.*, p.49, 2001.
13. F. Ducroquet *et al.*, *IEEE TED*, p.1816, 2000.
14. A. Yagishita *et al.*, *IEDM Tech. Dig.*, p.785, 1998.
15. Q. Lu *et al.*, *Symp. VLSI Tech.*, p.72, 2000.
16. H. Michaelson, *APL*, **48**, 4729, 1977.
17. I. Polishchuk *et al.*, *IEEE Elect. Device Letters* (in press).
18. P. Ranade *et al.*, *Electrochem. & Solid State Lett.*, **4**, 85, 2001.
19. Q. Lu *et al.*, *Symp. VLSI Tech.*, p.45, 2001.
20. P. Ranade *et al.*, to be published.