Reaction/annealing pathways for forming ultrathin silicon nitride films for composite oxide–nitride gate dielectrics with nitrided crystalline silicon–dielectric interfaces for application in advanced complementary metal–oxide–semiconductor devices

G. Lucovsky[a)]

Department of Physics, Electrical and Computer Engineering, and Materials Science and Engineering, North Carolina State University, Raleigh, North Carolina

Aggressive scaling of complementary metal–oxide–semiconductor (CMOS) devices requires gate dielectrics with an oxide equivalent thickness, $t_{ox,eq} \sim 1$ nm or less by the product introduction year 2012. Direct tunneling presents a significant performance limitation in field-effect transistors (FETs) with homogeneous oxide gate dielectrics <1.7 nm. Boron diffusion from p^+ poly-Si gate electrodes in p-channel FETs leads to additional electrical problems for oxide thicknesses <3 nm. Interfacial nitridation improves reliability in n-channel FETs; however, by itself, it is not effective in p-type metal–oxide–semiconductor FETs due to boron pileup at the Si–dielectric interface. Proposed solutions include top-oxide surface nitridation and the integration of composite oxide–nitride dielectrics into CMOS devices. This *review* discusses the integration of hydrogenated silicon nitride films, prepared by remote plasma-enhanced chemical-vapor deposition, into electrical devices with composite oxide–nitride (ON) gate dielectrics. FET devices with ON dielectrics having the same oxide-equivalent thickness, t_{ox-eq} and gate dielectric capacitance as devices with homogeneous oxide gate dielectrics display improved performance and reliability. However, reductions in direct tunneling current due to increased physical thickness are below expectations based on tunneling calculations which assume the tunneling mass of electrons in nitride films is approximately the same as in SiO$_2$. The combination of a lower electron tunneling mass and a reduced conduction-band offset energy (i) places important limitations on the extent to which devices with ON gate dielectrics can meet the aggressive scaling needed in advanced CMOS devices, and (ii) raises important questions that have to be addressed when evaluating alternative high-K dielectrics such as Ta$_2$O$_5$, TiO$_2$, and Al$_2$O$_3$. However, tunneling can be reduced by combining monolayer interface nitridation with ON stacks. © *1999 American Vacuum Society.* [S0734-2101(99)19604-0]

I. INTRODUCTION

As in-plane device dimensions of complementary metal–oxide–semiconductor (CMOS) devices in ultra-large-scale integrated circuits are aggressively scaled to <100 nm to achieve higher levels of speed and integration, there must also be decreases in the oxide-equivalent thickness of dielectrics, t_{ox-eq}, to <2 nm to maintain current levels required for fast circuit operation.[1] Direct tunneling current increases exponentially with decreasing thickness establishing an important limitation on the use of SiO$_2$ as a gate dielectric. For example, an oxide thickness of <3 nm defines the regime of *ultrathin oxides* in which direct tunneling is the dominant mechanism for current transport through these films. A practical limitation for homogeneous oxide dielectrics is $t_{ox-eq} \sim 1.7$ nm, the thickness at which the tunneling current at an oxide bias of ~ 1 V reaches a level of ~ 1 A cm^{-2}. The technology challenge is to increase gate dielectric physical thickness to reduce tunneling current while maintaining a t_{ox-eq} that corresponds to a significantly thinner SiO$_2$ film by using alternative insulators with dielectric constants higher than SiO$_2$. These alternative dielectrics include silicon ni-

tride, Si$_3$N$_4$, as well as other transition-metal binary and ternary oxides such as Ta$_2$O$_5$ and ZrSiO$_4$, respectively. These alternative dielectrics can be incorporated into stacked structures, such as the oxide–nitride (ON) composites, and be combined with monolayer-level nitrided Si–SiO$_2$ interfaces. This article focuses on the nitrided oxides, in particular, stacked ON structures and nitrided Si–SiO$_2$ interfaces, but does not address the other alternative high-K materials.

Selective incorporation of nitrogen (N) atoms into advanced gate dielectrics (i) reduces defect generation at the Si–SiO$_2$ interface when incorporated at concentrations from less than about 1 at. % ($\sim 2-7 \times 10^{12}$ cm^{-2}) to monolayer ($\sim 7-8 \times 10^{14}$ cm^{-2}) levels,[2–5] (ii) allows use of physically thicker stacked dielectrics when silicon nitride layers are incorporated into the body of gate dielectric as in ON stacks,[6,7] and (iii) reduces boron (B) atom penetration out of heavily doped p^+-polycrystalline silicon (poly-Si) gate electrodes through the dielectric films to the Si–SiO$_2$ interface when the nitride layers are at the polycrystalline Si–dielectric interface.[8–10] This review demonstrates separate and independent control of N-atom incorporation in these different parts of oxide gate dielectrics through combined use of low-temperature plasma-assisted processing at 300 °C, and low-
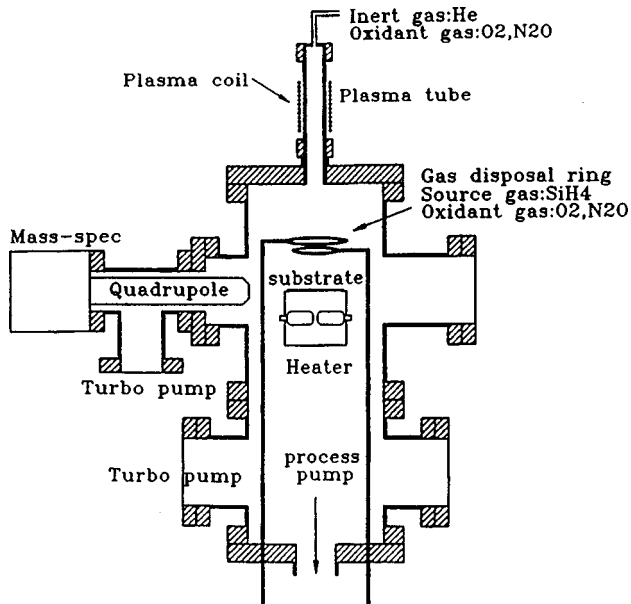

FIG. 1. Schematic diagram of remote plasma processing chamber.

thermal-budget rapid thermal annealing (RTA), e.g., 30 s at 900 °C.[2,5–8,10–12] This review is restricted to devices in which the nitride layers are prepared by remote plasma processing, and does not address formation of nitrided gate dielectrics by thermal processing, rapid thermal chemical-vapor deposition,[13] jet vapor deposition,[14] or other plasma deposition processes, where devices with similar performance have also been reported.

II. EXPERIMENTAL PROCEDURES

Dielectric layers for the device studies of this article have been prepared primarily by remote plasma-assisted processing, a technique that is differentiated from conventional or direct plasma processing in three ways: (i) it provides selective excitation of source and carrier gases as determined by their point of injection into the system, either through the plasma tube, or through the downstream showerhead injection ring; (ii) the deposition substrate is outside of the plasma glow region, and (iii) the source gases injected downstream from the plasma generation region are prevented from back-streaming into the plasma generation region by gas flow and process pressure (see Fig. 1).[15] The integration of remote plasma processing chambers into multichamber systems has made it possible (i) to interrupt plasma-assisted oxidation, nitridation and/or deposition processes and then, without removing the sample from an ultra-high-vacuum (UHV) compatible environment, perform on-line chemical analysis by Auger electron spectroscopy (AES) and (ii) to integrate on-line sequences of plasma processing with rapid thermal annealing.[4] Figure 2 shows differential AES data as a function of nitridation time. Figure 3 shows secondary ion mass spectrometry (SIMS) depth profile data corresponding to the nitridation conditions in Fig. 2. Figure 4 shows the integrated SIMS areal density of Fig. 3 plotted as a function of nitridation time. These data have been normalized to a concentra-
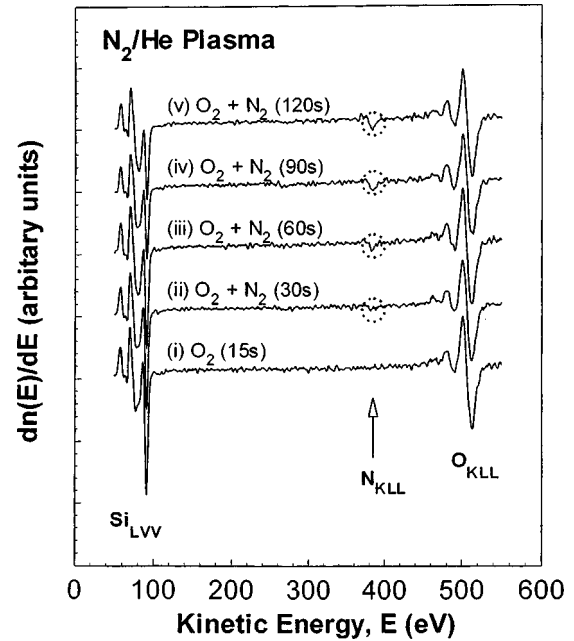

FIG. 2. Time evolution of differential AES spectra for the O_2 remote plasma-assisted oxidation process combined with postoxidation nitridation from a N_2/He plasma for different exposure times.

tion of ~1 ML ($7 \pm 1 \times 10^{14}$ cm^{-2}) for a 90 s exposure time. This concentration was determined from the SIMS data in Fig. 3 by using a nitride interface layer standard, and confirmed by nuclear reaction analysis. Localization of nitrogen atoms at the Si–SiO$_2$ interface was confirmed (i) by optical second-harmonic generation, (ii) by angle-resolved x-ray photoelectron spectroscopy (ARXPS), and (iii) by
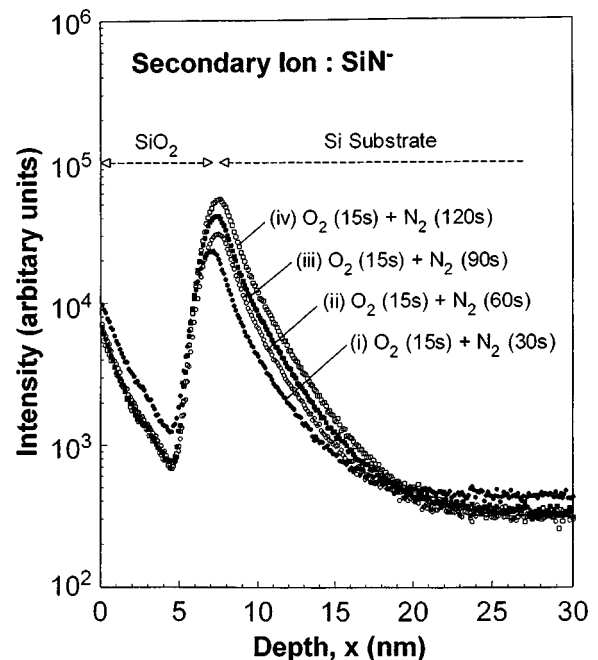

FIG. 3. Secondary ion mass spectrometry depth profiles for the N_2/He plasma nitridation process. The plasma processed interface has been overcoated with ~5 nm of plasma-deposited SiO$_2$.
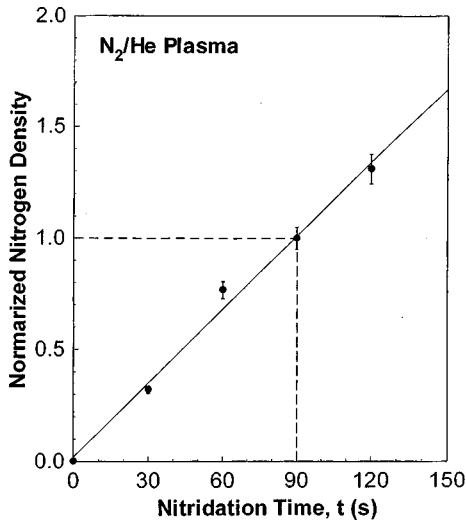
FIG. 4. Nitrogen concentration vs nitridation time. The value of one (1) corresponds to approximately ($\pm 10\%$) one monolayer of interfacial nitrogen atoms.

comparing in Fig. 5 the intensity of (a) the N signal with (b) the ratio of the N signal to the substrate Si signal, both as functions of time. A plot of the N signal intensity is sublinear due to the attenuation of AES electrons through the oxide film, whereas a plot of the N intensity, normalized to the Si substrate signal, which is also attenuated by passage through the oxide region, is linear, consistent with nonuniform incorporation.

Device fabrication has integrated different combinations of four 300 °C remote plasma-assisted processes in combination with postdeposition rapid thermal annealing: (i) plasma-assisted oxidation for Si–SiO$_2$ interface formation, (ii) plasma-assisted interface nitridation, (iii) deposition of "bulk" oxide and nitride films by remote plasma-enhanced chemical-vapor deposition (RPECVD), and (iv) plasma-assisted nitridation of oxide top surfaces (see Table I).[16,17]
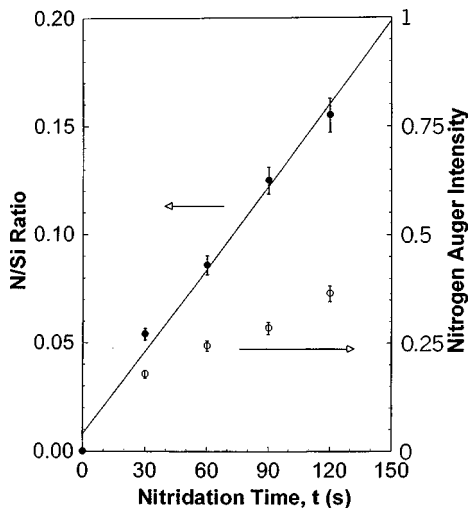

FIG. 5. Comparison of (i) the ratio of the amplitudes of the N$_{KLL}$ and substrate Si$_{LVV}$ feature peaks as a function of nitridation time and (ii) the amplitude of the N$_{KLL}$ feature peak as a function of nitridation time.

This article focuses primarily on the nitride layers in ON stacked dielectrics prepared by the processes labeled (iii) and (iv).

Device grade silicon nitride films have been prepared by RPECVD at 300 °C using either N$_2$ of HN$_3$ as the source gas for N, and SiH$_4$ as the source gas for Si. Previous publications have focused on the local atomic bonding of Si, N, and H in these films as a function of source gas mixtures, e.g., the ratio of NH$_3$ to SiH$_4$.[6,7,18,19] As-deposited films prepared from 10:1 NH$_3$/SiH$_4$ mixtures provided optimized performance of nitride films as gate dielectrics for thin-film transistors,[20] whereas annealing these films at 400–450 °C in forming gas, and using them as a constituent layer in stacked oxide–nitride–oxide (ONO) gate dielectrics yields marginal performance.[6,7] On the other hand, annealing the ONO stacks at 900 °C for approximately 30 s in an inert ambient, e.g., Ar, and then performing a 400–450 °C forming gas anneal results in excellent device performance with reliability generally exceeding oxide dielectrics with the same oxide-equivalent thickness.[6,7,21] These differences in performance between (TFTs) and field-effect transistors (FETs) with respect to processing temperatures have been resolved by employing AES and infrared spectroscopy for chemical characterization.[20] The optimized as-deposited films display Si–N bonding by on-line AES, whereas the films deposited with lower flow rates of NH$_3$ or N$_2$ to SiH$_4$ are subnitrides with both Si–Si and Si–N bonds detected by AES. Bonded hydrogen occurs predominantly as Si–H in the subnitrides, whereas SiN–H groups dominate in the hydrogenated nitrides. Typical bonded hydrogen concentrations in the as-deposited films are ~20 at. % for films grown from N$_2$, and ~25–30 at. % for films grown from NH$_3$.[17,18] Optimized films display predominantly SiN–H bonding, but there are IR-detectable Si–H groups as well. Upon annealing at temperatures greater than the 300 °C deposition temperature, there is a loss of bonded hydrogen with an activation energy of ~0.4 eV. For annealing at temperatures greater than 500 °C, loss of bonded hydrogen continues, but additional absorption begins to appear in the Si–N bond stretching regime (see Fig. 6). Films annealed to 900 °C, and thereby optimized for gate dielectrics in MOS devices, typically display bonded hydrogen concentrations primarily in *isolated* SiN–H groups (i.e., not having other SiN–H groups as nearest neighbors) of 10–15 at. %, but show no detectable IR absorption in Si–H configurations. These isolated SiH–H bonds are stable up to annealing temperatures of at least 1200 °C.[18]

Top surface nitridation has also been accomplished by a 300 °C plasma-assisted process in which the process pressure is reduced to 100 mTorr effectively changing the plasma reactor from a remote to an after-glow mode.[8,9] In the after-glow mode the concentration of charged N species at the film surface is many orders of magnitude higher than it is in the remote reactor process regime. This after-glow process is used to form a nitride layer on the top surface of a previously deposited oxide dielectric. Figure 7 indicates the intensity of the N$_{KLL}$ AES feature as a function of plasma exposure time

TABLE I. Process steps for forming device-quality stacked gate oxide dielectrics.

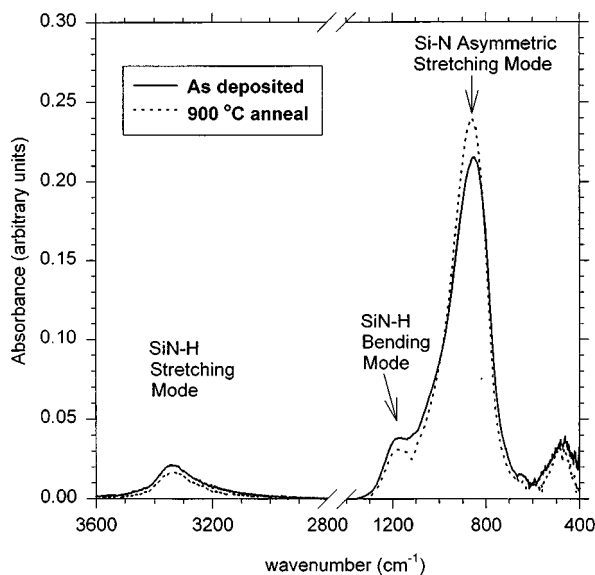| Process step	Process conditions	Processing results
(a) Remote plasma-assisted oxidation	Substrate temperature 300 °C Process pressure: 300 mTorr Plasma excited mixture: He/O$_2$ (200 sccm He, 20 sccm O$_2$) Time: ~15–30 s	*In situ* substrate cleaning (reduces C and F level) Forms Si–SiO$_2$ interface Growth passivating oxide: ~0.5 nm Introduces suboxide bonding at Si–SiO$_2$ interface
(b) Remote plasma-enhanced CVD	Substrate temperature 300 °C Process pressure: 300 mTorr Plasma excited mixture: He/O$_2$ (200 sccm He, 20 sccm O$_2$) Downstream mixture: He/SiH$_4$ (20 sccm He/0.4 sccm SiH$_4$	Forms body of dielectric film Deposition rate: 2.5–5.0 nm/min Stoichiometric SiO$_2$ No IR detectable Si–H or Si–OH Low Si–OH (≪5 at. % H)
(c) Remote plasma-assisted interface nitridation	Substrate temperature 300 °C Process pressure: 300 mTorr Plasma excited mixture: He/N$_2$ (160 sccm He, 60 sccm N$_2$) Time: ~90 s	Inserts approximately one monolayer of nitrogen atoms at Si–SiO$_2$ interface Nitrogen is localized at interface Nitrogen concentration scales with time
(d) Remote plasma-assisted top surface nitridation	Substrate temperature 300 °C Process pressure: 100 mTorr Plasma excited mixture: He/N$_2$ (200 sccm He, 20 sccm N$_2$) Time: ~10–20 min	Forms ~2 molecular layers silicon nitride at top surface of a plasma or thermally grown oxide sufficient to suppress boron diffusion
(e) Rapid thermal annealing	Temperature 900 °C Time: ~30 s Low pressure or atmospheric inert gas ambient (e.g., Ar)	Reduces oxidation-induced suboxide bonding at Si–SiO$_2$ interface Promotes densification of oxide films Reduces bonded H (mostly in nitrides)

FIG. 6. Infrared absorption spectrum for plasma-deposited hydrogenated silicon nitride film: (i) as deposited at 300 °C, and (ii) after a postdeposition 900 °C rapid thermal anneal.
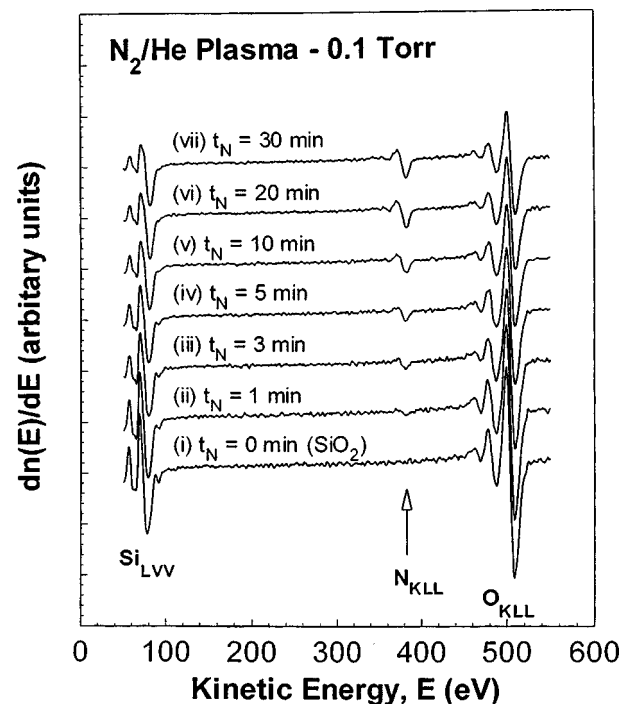

FIG. 7. Differential AES spectra for remote plasma-assisted top oxide surface nitridation.
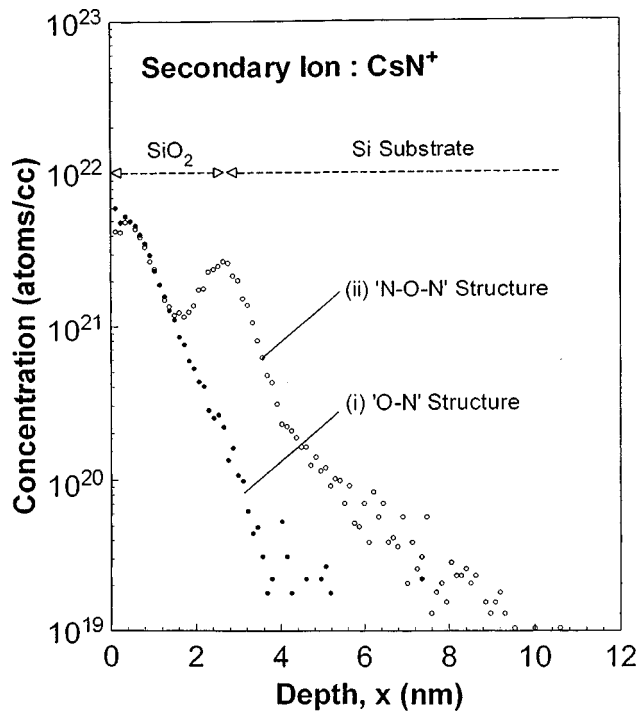
FIG. 8. Secondary ion mass spectrometry depth profiles for (i) the top-surface N_2/He plasma nitridation process, and (ii) the combined top-surface and interface N_2/He plasma nitridation processes.

plasma excitation of He and downstream injection of N_2. ARXPS has demonstrated that nitridation is restricted to the top surface of the oxide film;[22] this has been confirmed by SIMS, which is shown in Fig. 8.

Finally, device fabrication procedures incorporated conventional photolithography, masking, ion implantation, dopant activation anneals, and postmetallization anneals.

III. RESULTS

This section discusses electrical measurements of metal–oxide–semiconductor (MOS) capacitors and FETs, and is organized into three parts which (i) demonstrate tunneling current reductions due to interface nitridation; (ii) quantifies the stopping of boron out-diffusion from p^+ polycrystalline Si; and (iii) discusses the performance of n-type MOS (NMOS) and p-type MOS (PMOS) FETs with stacked ON gate dielectrics and nitrided Si–SiO$_2$. All device structures were subjected to 30 s, 900 °C rapid thermal anneal prior to the deposition of either Al or polycrystalline Si gate electrodes. This section also includes a short discussion of differences in performance between devices that include stacked ON dielectrics and devices with Si$_3$N$_4$ dielectrics alone.

A. Interface nitridation in MOS capacitors

Figures 9(a), 9(b), and 9(c) present current–voltage characteristics for NMOS and PMOS structures with and without intentionally nitrided interfaces.[5] It is important to note that this interface nitridation process is run at 300 mTorr, a pressure in which the plasma is confined to the plasma generation region, and at which the interface nitridation process

utilizes neutral species, such as N atoms and excited nitrogen molecules, N_2^*. As noted above, the top-surface nitridation is qualitatively different and is run at 100 mTorr and utilizes charged nitrogen species instead of neutrals. The plots in Fig. 9(a) indicate the reductions in tunneling current increase as interface nitridation is increased with an apparent saturation of the effect occurring for monolayer coverage. The plots in Figs. 9(b) and 9(c) are for monolayer interface nitridation. Capacitance–voltage (C–V) characterizations for devices in Fig. 9 indicate that oxide-equivalent capacitance is essentially the same for each pair of devices in Figs. 9(b) and 9(c), and the group of devices in Fig. 9(a). This means that the pairs of devices in Figs. 9(b) and 9(c) and the groups of devices in Fig. 9(a), respectively, have essentially the same thickness to ±0.1 nm. The effect of interface nitridation is clearly to reduce tunneling. It is important to note that the fractional reduction of tunneling current (i) saturates at a level of monolayer coverage; (ii) is independent of the injection direction, substrate, or gate electrode, (iii) is independent for the gate electrode material; but (iv) is not independent of oxide thickness, showing different behavior in the Fowler–Nordheim [Fig. 9(a)] and direct tunneling [Figs. 9(b) and 9(c)] regimes. These data are consistent with a tunneling electron mass of $\sim 0.5 m_0$, and a conduction-band offset energy of ~ 3.15 eV. The reduction in Fowler–Nordheim tunneling is about a factor of 10 gate injection, and increases to about 50–60 for substrate injection. In other publications, it is demonstrated that the reductions in tunneling current are associated with changes in suboxide interfacial bonding that accompanies the interface nitridation.[23]

B. Suppression of boron atom transport for PMOS devices

The minimum thickness of silicon nitride in contact with a p^+-polycrystalline gate electrode that is needed to stop boron out-diffusion during a high-temperature dopant activation anneal (900–1050 °C) has been estimated to be at most 0.8±0.1 nm, or approximately 2 molecular layers, corresponding to a nitrogen atom areal density of $4.5 \times 0.5 \times 10^{15}$ cm^{-2}.[12] The areal density of nitrogen atoms is determined from the physical density of the nitride material (~ 3.1 gm cm^{-1}) and mass of one molecular unit of Si$_3$N$_4$ (140 gm). Figure 10 includes C–V traces for capacitors with plasma-deposited and annealed nitride layers of either 0.4±0.1 or 0.8±0.1 nm, and for a control device with oxide dielectric without a nitride capping layer. The effect of boron transport out of the p^+-poly-Si is to shift the C–V characteristic to more *positive values* increasing the flatband and threshold voltages. The C–V characteristics for devices with top nitride films thicker than 0.8 nm are essentially the same as those for nitride films that are 0.8 nm thick. It is significant to note that 0.4 nm of nitride yields a C–V *position* intermediate between that of a dielectric with a 0.8 nm nitride layer, and the oxide dielectric indicating that some boron penetration and transport has occurred through the 0.4 nm film.

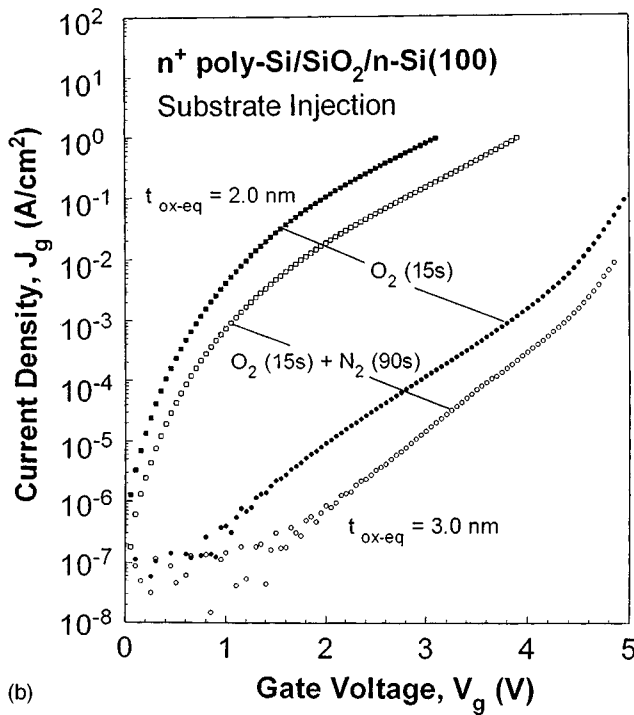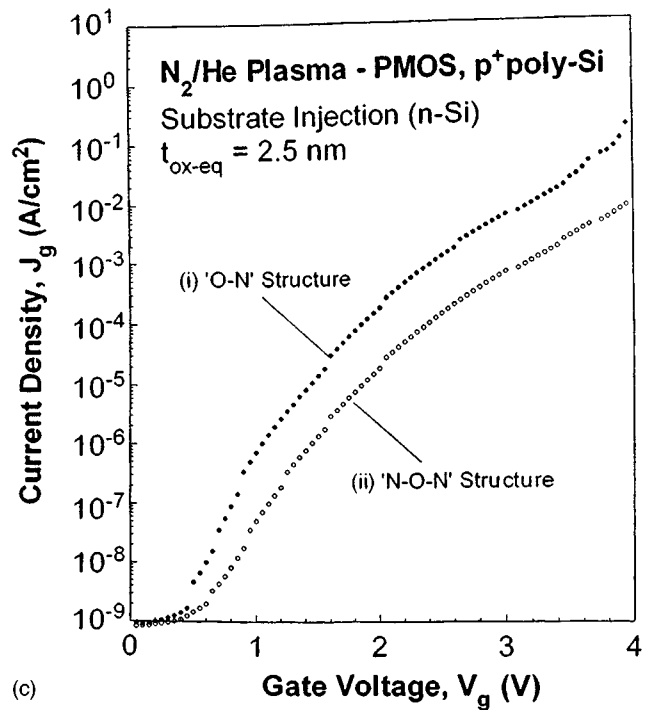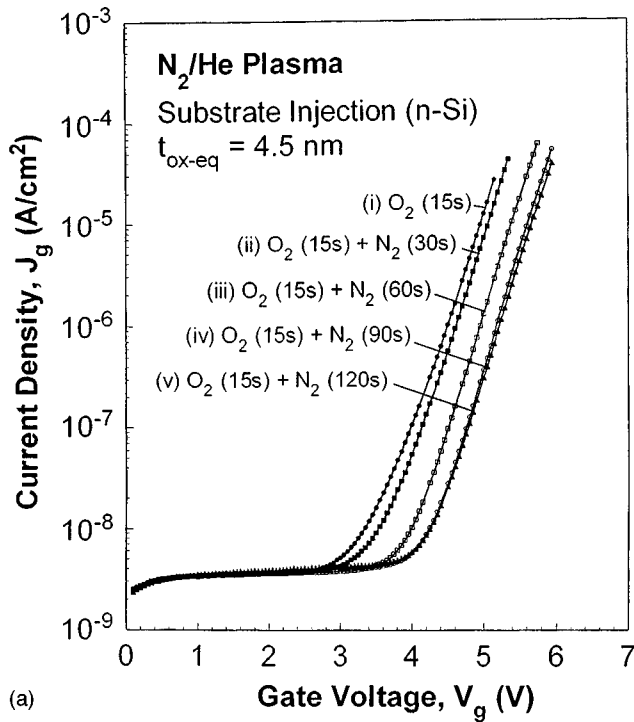Figures 11(a) and 11(b) indicate a series of experiments

FIG. 9. Current density/oxide voltage traces for capacitors with: (a) t_{ox} ~ 4.5 nm, and (b) $t_{ox} \sim 3$ and 2, and (c) $t_{ox} \sim 2.5$ nm.

that provide insight into the boron blocking mechanism of the 0.8 nm nitride films. The traces in Fig. 11(a) are for PMOS devices in which the thickness of the top nitride or oxynitride layer has been fixed at 0.8 nm: (i) a 0.8 nm nitride film; (ii) two films with 0.8 nm thick oxynitride alloy films with compositions of 30% and 70% SiO_2, respectively; and (iii) a control oxide. A similar set of traces is included in Fig. 11(b), where instead of top-layer thickness being held constant, the top-layer nitride atom areal density was held a level corresponding to 0.8 nm of nitride ($\sim 4.5 \pm 0.5 \times 10^{15}$ cm^{-2}).

The effects of boron penetration are referenced to the oxide layers. Keeping the areal density fixed, and increasing oxynitride alloy thickness is more effective at blocking boron transport than keeping the top-layer thickness constant and increasing the SiO_2 alloy fraction. Studies made on n-type substrates, but using Al, rather than p^+-poly-Si gate electrodes, indicated that fixed charge levels were at most in the low 10^{11} cm^{-2} regime. As such, shifts in flatband voltage, relative to calculated values, were significantly smaller than the shifts in flatband voltage due to boron penetration to the
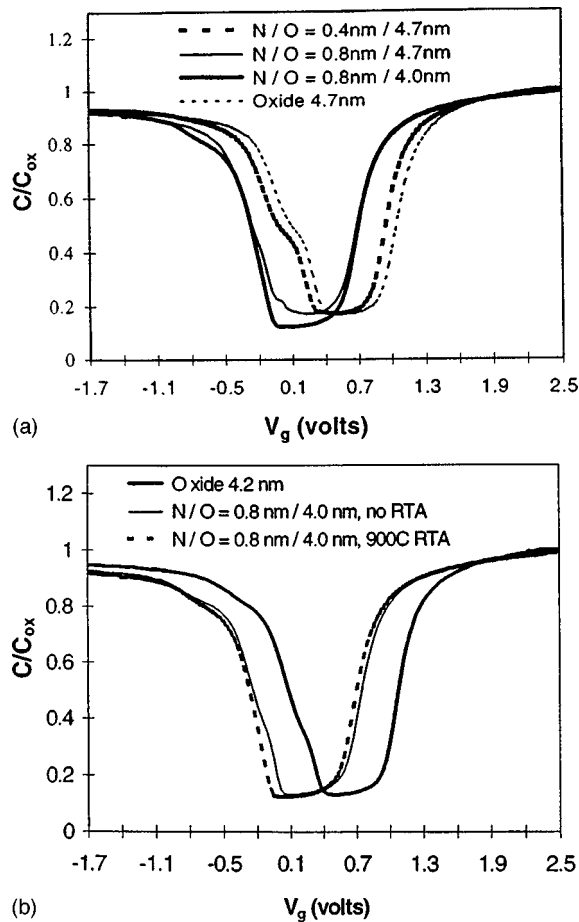
FIG. 10. Normalized quasistatic $C-V$ curves for thermal oxide, and 0.4 and 0.8 nm top nitride deposited onto thermal oxides. Curves are shifted due to boron penetration through thin gate material. The dopant activation annealing is 1000 °C for 60 s.
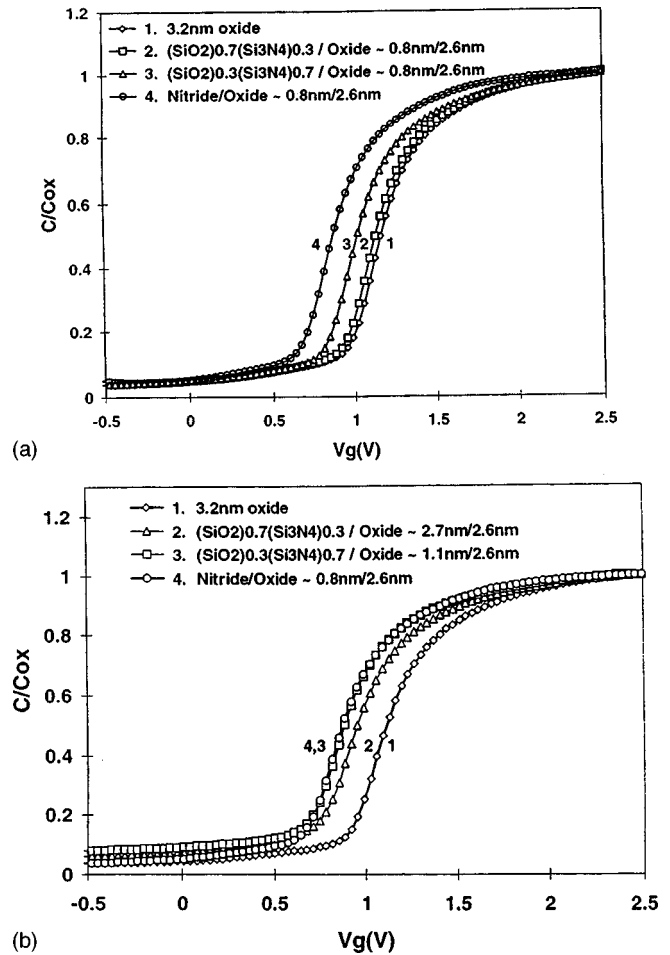

FIG. 11. $C-V$ traces for different top-surface nitride barrier layers. Each set of traces has an oxide dielectric for reference (No. 1) in which the flatband voltage has been shifted to positive values by B-atom penetration to the Si–SiO$_2$ interface. Each set of traces also includes a 0.8 nm nitride layer (No. 2) which effectively completely suppresses B-atom transport. The traces in (a) are for a fixed barrier layer thickness of 0.8 nm, and those in (b) are for a fixed area density of nitrogen atoms, $\sim 4.5 \times 10^{15}$ cm^{-2}.

Si–SiO$_2$ interface. Additionally, shifts in flatband voltage due to fixed positive charge are in the *opposite direction* to shifts associated with boron penetration to the Si–SiO$_2$ interface, so that these would tend to decrease the quantitative values of the flatband voltage shifts in Figs. 11(a) and 11(b).

C. Properties of NMOS and PMOS FETs

Plasma-deposited nitride layers have been incorporated into NMOS and PMOS FETs and have been discussed at some length in Refs. 11 and 12, respectively. The oxide layers employed were formed either by thermal oxidation in O$_2$, or plasma-assisted oxidation using O$_2$ as the source gas. Weak interface nitridation at the $\sim 2-5 \times 10^{12}$ cm^{-2} level occurred during the nitride deposition, and as demonstrated in Ref. 12, this was sufficient to improve device reliability with respect to devices with oxide dielectrics alone. NMOS and PMOS FETs were made with different oxide and nitride layer thicknesses, and with $t_{\text{ox-eq}}$ values ranging from about 1.8 to 3.5 nm. In all instances the following results were obtained: (i) transistor drive currents were determined by $t_{\text{ox-eq}}$, or equivalently, the capacitance of the dielectric layer; (ii) reliability was improved with respect to oxide dielectrics

with the same $t_{\text{ox-eq}}$; and finally (iii) direct tunneling leakage was reduced relative to oxide dielectrics with the same $t_{\text{ox-eq}}$. These electrical properties of PMOS FETs are illustrated, respectively, in Figs. 12(a), 12(b), and 12(c) and Fig. 13. Figure 13 displays direct tunneling in PMOS devices with $t_{\text{ox-eq}} \sim 1.9$ nm, and with different nitride-to-oxide thickness, t_n/t_o, ratios. The reduction in tunneling current in the two devices with ON gate dielectrics is about the same even though the t_n/t_o ratios, and physical thicknesses, are markedly different.

PMOS devices have recently been fabricated with interface nitridation at the $7-8 \times 10^{14}$ cm^{-2} level, and with $t_{\text{ox-eq}} \sim 1.6$ nm.[25] The drive currents scale with the increased capacitance, and the reliability shows improvements similar to what has been discussed above. However, significant decreases in the tunneling current are obtained, reflecting separate and independent contributions from (i) interface nitridation and (ii) the increased physical thickness of the ON stack
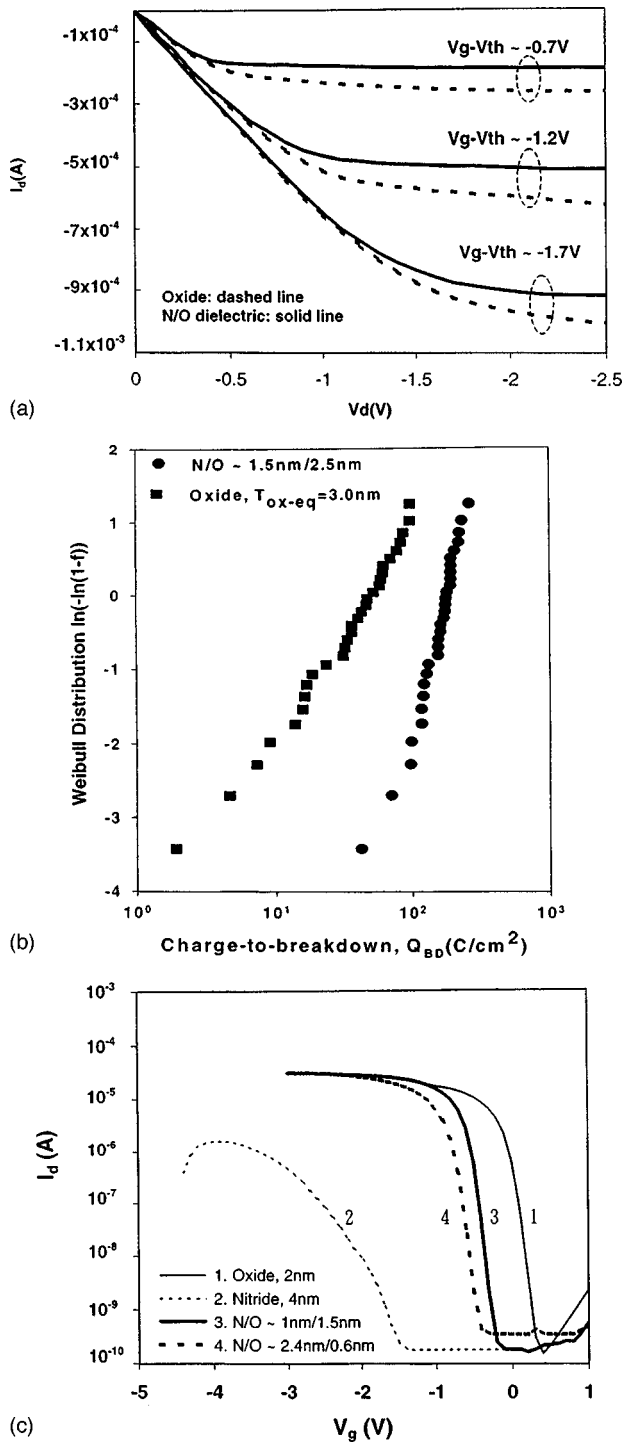
(a)

(b)

(c)

FIG. 12. Properties of PMOS FETs with ON gate dielectrics. Comparisons are made with a FET with a control oxide: (a) drain current as a function of drain voltage for different normalized gate voltages, (b) charge to breakdown, and (c) substrate injection tunneling current.
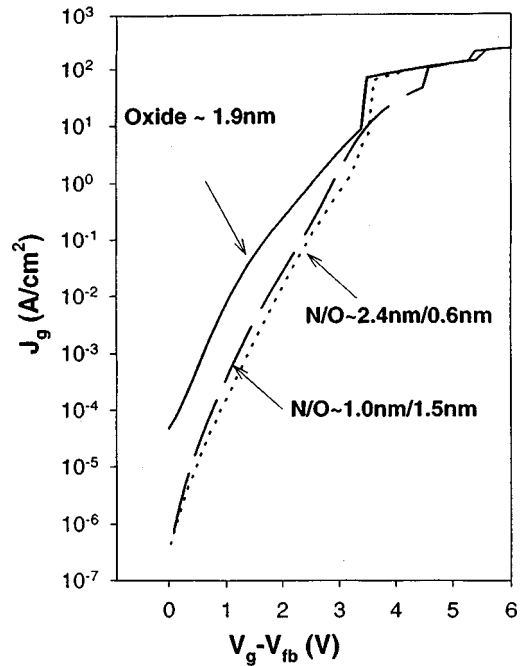

FIG. 13. (a) $C-V$ characteristics demonstrated shift in flatband voltage due to positive charge, and increased separation between high-frequency and quasistatic plots due to interface trapping accompanying direct deposition of thin nitride films onto Si. (b) I_d-V_g characteristics for PMOS FETs with (i) a 4 nm nitride layer, (ii) a 0.6 nm oxide separating a 2.4 nm nitride from the Si substrate, and (iii) a 1.5 nm oxide separating a 1.0 nm nitride from the Si substrate. The threshold voltage shift between (ii) and (iii) is due in part to substrate doping differences (0.16 V) and in part to positive charge at the oxide–nitride interface (0.04 eV).

as compared to an oxide with the same t_{ox-eq}. These devices are discussed in more detail in Ref. 25.

D. Nitrided interfaces

Devices made with nitrided interfaces have been discussed at length in Ref. 24. PMOS devices with single-layer

optimized nitride gate dielectrics perform poorly showing threshold voltage shifts of about 1 V and saturated mobility degradations of about 50, whereas NMOS devices show much reduced threshold voltage shifts, and mobility degradations of about 2. The mobility degradations are referenced to devices with Si–SiO$_2$ interfaces, and sufficient nitriding of the top surface of the oxide to suppress boron out-diffusion during dopant activation.

IV. DISCUSSION

A. Control of nitrogen profiles

The experimental data presented above have shown that remote plasma-assisted processing combining (i) plasma-assisted oxidation, (ii) plasma-assisted interface and top-surface nitridation, and (iii) plasma-assisted film deposition can control nitrogen profiles in ultrathin gate dielectrics with values of t_{ox-eq} extending to at least 1.8 nm. The nitrogen concentrations have been determined by SIMS, AES, and nuclear reaction analysis (NRA). Localization of N atoms at the Si–SiO$_2$ interface has been determined by AES, and confirmed by angular-resolved x-ray photoelectron spectroscopy,[15] and optical second-harmonic generation.[17]

B. Si–SiO$_2$ interface nitridation

Si–SiO$_2$ interface nitridation is important in the formation of ultrathin stacked oxide–nitride gate dielectrics. The ex-

TABLE II. Average bonding coordination at Si–dielectric interfaces.

Material system	Average coordination (N_{av})	Electrical quality
Si–SiO_2 (1.5 molecular layers)	2.8	Excellent
Si–Si_3N_4 (1.5 molecular layers)	3.5	Very poor (Ref. 24)
Si–{SiO_2}(t)–Si_3N_4	t=0.6 nm: 3.0	Very good (Ref. 30)
t=oxide layer thickness	t=1.5 nm: 2.9	Excellent (Ref. 30)
Si–{Si_3N_4}(t)–SiO_2	t=0.4 nm: 3.3	Poor (Ref. 31)
t=oxide layer thickness	t=0.8 nm: 3.4	Poor (Ref. 31)
Si–N–SiO_2 {1 monolayer (ML)}	2.8	Excellent (Ref. 17)
Si–$(SiO_2)_{0.977}${Si_3N_4}$_{0.023}$	2.3 at. % N:2.8	Excellent (Ref. 40)
Si–$(SiO_2)_{0.89}${Si_3N_4}$_{0.11}$	11 at. % N:3.0	Poor (Ref. 40)
Si–TiO_2}[a] (1.5 molecular layers)	4.0	Unreported
Si–Ta_2O_5}[b] (1.5 molecular layers)	3.5	Unreported
Si–Al_2O_3}[b] (1.5 molecular layers)	3.6	Unreported

[a]Average coordination: (Ti)=6, (O)=3.0 (rutile/anatase bonding).
[b]Average coordination: (Ta)=6, (O)=2.4 (Ref. 41).
[c]Average coordination: Al=(4.5), (O)=3.0 [3:1 ratio of tetrahedral to octahedral sites (Ref. 42)].

perimental data in Refs. 5 and 17 and displayed in Figs. 9 and 12 indicate that interface nitridation reduces tunneling currents and improves device reliability as well.

C. Improvements in NMOS and PMOS devices with stacked oxide/nitride gate dielectrics

Improvements in NMOS and PMOS FETs with ultrathin stacked oxide–nitride gate dielectrics have been identified in Refs. 11 and 12. As noted above, these devices have low levels of interface nitridation that occur during the nitride deposition step (~1% of the surface Si atoms are nitrided). The primary improvements in these devices with stacked gate dielectrics as compared to devices with oxide dielectrics at the same $t_{ox\text{-}eq}$ are (i) improved reliability, (ii) improved suppression of boron out-diffusion from p^+-poly-Si gate electrodes in PMOS devices, and (iii) tunneling current reductions of about 7–10 [see Figs. 12(a), 12(b), and 12(c)]. The tunneling currents are well below expectations and are sufficiently small as to raise serious questions regarding implementation of devices with stacked oxide–nitride gate dielectrics into commercial devices. However, significant improvements have been made in tunneling current reductions by combining ON stacks with heavily nitrided interfaces; for example, reductions of tunneling current in excess of two orders of magnitude have reported in PMOS FETs with $t_{ox\text{-}eq}$~1.6 nm.[25]

D. Limitations on interface nitridation

Experiments described in Ref. 24 have established that devices with nitride gate dielectrics do not perform as well as stacked oxide/nitride devices with interfacial oxide layers as thin as 0.5–0.6 nm. The performance of PMOS devices with nitride gate dielectrics is significantly poorer than NMOS devices indicating that defects at the Si–Si_3N_4 interface device from states in the lower half of the band gap that are positively charged when the Fermi level resides close to the valence-band edge (see Figs. 13).[24] Although these states have not been identified by any spectroscopic technique, it is

likely that they are Si-atom dangling bonds that are not compensated by hydrogen due to steriochemical constraints.

The significant difference in performance between devices with nitride gate dielectrics and those in which superficially thin SiO_2 layers have been interposed between the Si substrate and the Si_3N_4 gate dielectric has been explained through an application of constraint theory to crystalline semiconductor dielectric interfaces.[26] This theory, originally developed for bulk glasses,[27] and recently extended to thin films,[20] has been further extended to the interface structures of this article. Table II includes some of the results presented in Ref. 26, where it has been shown than an average interface bonding coordination of N_{av}~3 separates defect interfaces of device-quality interfaces. For example, Table II shows that $N_{av} \leqslant 3$ for Si–SiO_2, monolayer nitrided Si–SiO_2, Si–SiO_2 (0.6) nm–Si_3N_4, but $N_{av} > 3$ for Si–Si_3N_4.

E. Limitations on ON stacks due to direct tunneling

Model calculations assuming equal electron tunneling masses in SiO_2 and Si_3N_4 have demonstrated tunneling currents decrease significantly with increases in physical thickness at constant $t_{ox\text{-}eq}$, as obtained by replacing SiO_2 with Si_3N_4 (k_n~7.6) or stacked oxide/nitride dielectrics.[28,29] For direct replacement of SiO_2 by Si_3N_4, a first-order estimate for tunneling reductions is obtained by applying the WKB approximation to a symmetric tunneling barrier. For small oxide voltages, the tunneling probability is proportional to $\exp(-2t/\lambda_{deB})$, where t is the film thickness, and λ_{deB} is the deBroglie wavelength for electrons in the dielectric $=(\hbar)/\{2m_i^* E_{bi}\}^{0.5}$; E_{bi} is the barrier height (conduction-band offset energy) and m_i^* is the effective electron tunneling mass, $i=o,n$. For Si_3N_4 substitutions to reduce direct tunneling, $t_n/\lambda_{deB}]n > t_o/\lambda_{deB}]_o$. t_n and t_o are the respective nitride and oxide layer thicknesses, $t_n = rt_o$, and r is the dielectric constant ratio, k_n/k_o~2, so that

$$r^2(m_n^* E_{bn}) > (m_o^* E_{bo}). \qquad (1)$$

Conduction-band offset energies at Si–SiO_2 and Si–Si_3N_4 interfaces are, respectively, ~3.15 and ~2.15 eV, so that

$m_n^*/m_o^* > 0.37$. m_o^* is $\sim 0.5m_o$, where m_o is the free-electron mass, so that tunneling current is reduced if $m_n^* > 0.18m_o$.[29] For $t_{ox-eq} \sim 2$ nm, reductions $> 10^3$ require $m_n^* \sim m_o^*$, or $0.5m_o$. Comparisons with model calculations demonstrate that this criterion also applies at larger applied bias voltages.[29]

Electrical evaluation of FETs with Si–Si$_3$N$_4$ interfaces deposited by remote plasma processing[24,30] have yielded densities of interface traps and fixed charge about three orders of magnitude higher than at Si–SiO$_2$, so direct replacement of SiO$_2$ by Si$_3$N$_4$ is not viable. However, interface defect densities could be reduced to Si–SiO$_2$ levels (mid-10^{10} cm^{-2}) by interposing ~ 0.6 nm of plasma-grown SiO$_2$ between the Si substrate and Si$_3$N$_4$ film forming a stacked ON gate dielectric.[24,30] Other studies prepared device-quality nitride films by jet vapor deposition[14] and rapid thermal chemical-vapor deposition (RTCVD).[13] These films typically require postdeposition anneals in oxidizing ambients forming thin SiO$_2$ layers at the Si–dielectric interface, and incorporating oxygen in the bulk dielectric.[13]

Using the WKB approximation, the condition for reduced tunneling through stacked ON dielectric is given by

$$(t_{ox-eq})\{2m_o^*E_o\}^{0.5} < t_o\{2m_o^*E_{bo}\}^{0.5} + t_n\{2m_n^*E_{bn}\}^{0.5}. \quad (2)$$

Since $t_{ox-eq} = t_o + t_n/r$, Eq. (2) reduces to the same condition as above, $m_n^* > 0.18m_o$. This constraint on m_n^* is supported by experiments on ON devices with $t_{ox-eq} \sim 1.9$ nm, and with physical thickness ranging from 1.9 to ~ 3 nm.[11,12] These devices display tunneling reductions of 8 ± 2 with respect to devices with SiO$_2$ dielectrics independent of the ratio of t_n/t_o which was varied from ~ 0.67 to 4. The reductions are consistent with $m_n^* \sim 0.3m_o$ as first estimated on the basis of ON devices with $t_n/t_o \sim 1$.[3,11] Similar reductions in direct tunneling for stacked ON dielectrics have been reported in Ref. 13, where the nitride film was deposited by RTCVD and subjected to a postdeposition anneal in an oxidizing ambient. These results have raised questions as to whether the many order of magnitude tunneling current reductions required for aggressively scaled devices can be realized with composite ON gate dielectrics in which a separate interface nitridation was performed. Nitrogen atoms diffuse to the Si–SiO$_2$ interface during the plasma-assisted deposition of the nitride films, but the areal density of nitrogen atoms at these interfaces are of the order of at most 1 at.% ($\sim 2-5 \times 10^{12}$ cm^{-2}).

As noted above combining ON stacks with a separate interface nitridation step that incorporates approximately 1 ML of interface nitrogen, tunneling currents in the ON stacks can be reduced by more than one order of magnitude.[25] The reduction of direct tunneling current in these devices then derives from two effects: (i) interface nitridation which gives about a factor of 10 [see Figs. 9 and 12(c)],[12] and (ii) physically thicker films which also contribute about a factor of 10. Alternatively, the reduction in tunneling current in the ON device may be due to monolayer-level interface nitridation which occurs during plasma nitride deposition and annealing. Other studies have demonstrated that plasma-assisted

nitridation of Si–SiO$_2$ interfaces at monolayer levels reduces both direct and Fowler-Nordheim tunneling up to approximately a factor of 50 with respect to devices with non-nitrided Si–SiO$_2$ interfaces for t_{ox-eq} from ~ 2 to 5 nm.[17,31] The dielectrics in these devices were plasma-deposited SiO$_2$, and the plasma nitridation process confined nitrogen to Si–SiO$_2$ interface.

A similar analytical approach has also been applied to other insulators with increased dielectric constant ratios. Recent calculations have shown that barrier heights between the conduction band of Si and many alternative high-K oxides is < 1.5 eV.[32] For example, the conduction-band offset energy between Si and Ta$_2$O$_5$ is estimated to be ~ 0.36 eV; however, interposition of a thin SiO$_2$ layer between the Si and Ta$_2$O$_5$ is likely to increase this relative band offset energy by about 0.5 eV. Assuming a nominal value of 25 for $k_{Ta_2O_5}$, the tunneling mass of electrons in Ta$_2$O$_5$ must be $> 0.1m_o^*$ or $\sim 0.05m_0$ for reduced direct tunneling, a condition which should be realized. Based on electron transport through/over Schottky barriers,[33] the relatively small barrier height between Si and Ta$_2$O$_5$, means that tunneling transport may not be the dominant leakage mechanism at bias voltages and interface fields used in advanced FET devices. The barrier height of Ref. 33 (including increases to interposition of the SiO$_2$ layer) is sufficiently low so that field-assisted thermionic emission over the Si/Ta$_2$O$_5$ barrier or thermally assisted tunneling through that barrier will promote electron injection directly into the band tail or conduction-band states of Ta$_2$O$_5$. This is supported by experiments which show that the electron current from Si substrates through Ta$_2$O$_5$ films is temperature activated and limited by a bulk transport process.[34,35] The activation energy for transport in Ref. 34 is close to the conduction-band offset energy calculated in Ref. 32, consistent with barrier-limited injection. Alternatively, other papers have suggested the bulk transport in Ta$_2$O$_5$ as a Poole–Frenkel trap controlled process giving a different interpretation to the activation energy.[32] If Poole–Frenkel conduction were the bulk transport process, then the Si–SiO$_2$ interfacial barrier layer in the devices of Ref. 5 would act as a Ohmic contact, supplying carrier flow-limited bulk transport requirements.[36]

Experiments indicate that the temperature-dependent leakage current in Ta$_2$O$_5$ devices is lower by several orders of magnitude than the tunneling current in SiO$_2$ for the same t_{ox-eq} extending to below 2 nm.[34] However, to qualify as a replacement for SiO$_2$, at least four criteria must be realized: (i) the gate dielectric must have a capacitance that corresponds to ~ 1 nm of SiO$_2$, (ii) the leakage current for a device with $t_{ox-eq} \sim 1$ nm must be at least three orders of magnitude less than that of an ~ 1 nm SiO$_2$ film, (iii) the defect generation rate under an appropriate stress bias test must correspond to a device lifetime of at least 10 years, and (iv) electron and hole channel mobilities in FETs must be degraded by no more than 5%–10% with respect to their values at Si–SiO$_2$ interfaces. Initial results on electron mobilities in n-channel devices with Ta$_2$O$_5$ and TiO$_2$ dielectrics and SiO$_2$

interfacial layers indicate significant channel mobility reductions.[34,37] Additional experimentation is needed to determine if these reductions are intrinsic to the materials in the gate stack, or if they are associated with particular processing conditions.

V. CONCLUSIONS

The results presented above have demonstrated that stacked ON gate dielectrics with nitrided interfaces can play an important role in meeting SIA Technology Roadmap goals.[1] For example, stacked ON gate dielectrics possess all of the potential advantages for gate dielectric nitridation; (i) interface nitridation improves performance and reliability, (ii) bulk nitride incorporation allows for increased physical thickness without increases in oxide-equivalent thickness and decrease in capacitance, and (iii) top-surface nitride layers block B-atom transport out of boron-doped p^+-polycrystalline silicon gate electrodes. Preliminary studies have shown that ON gate dielectrics are compatible with elemental and compound metal gate electrodes such as WN_x for PMOS.[38]

Based on the research reported to date, stacked ON structures with physical thicknesses of 2.5 nm, and $t_{ox-eq} \sim 1.5$ nm should meed the SIA Technology Roadmap goals projected to the 2006–the 2009 time frame.[1] However, recent research results from IBM, which have been reported at IEDM, raise some important questions on continued scalability that apply not only to the ON composite dielectrics of this article, but also to other high-K alternative dielectrics.[39] The IBM study indicates that defect generation falls off too slowly with decreasing bias voltage for the tunneling limitations of the 1997 SIA Technology Roadmap to apply to devices which require an extrapolated lifetime of more than 10 years. This can place a severe restriction on the implementation of CMOS devices with ON gate stacks. The tunneling current can easily be lower than the 1 A cm^{-2} limitation for $t_{ox-eq} \sim 1.7$ nm, and perhaps to also low as 1.5 nm; however, this may not to be adequate if the criterion of Ref. 39 is applied. Additional studies are clearly necessary to redefine tunnel and other leakage current limitations for aggressively scaled CMOS devices, particularly in the area of their impact on device reliability.

ACKNOWLEDGMENTS

[1] The SIA Roadmap for Semiconductor Technology, 1997, Santa Clara, CA.

[2] D. R. Lee, G. Lucovsky, M. R. Denker, and C. Magee, J. Vac. Sci. Technol. A **13**, 607 (1995); D. R. Lee, C. G. Parker, J. R. Hauser, and G. Lucovsky, J. Vac. Sci. Technol. B **13**, 1778 (1995).

[3] M. L. Green, D. Brasen, K. W. Evens-Lutterodt, L. C. Feldman, K. Krisch, W. Lennard, H. T. Tang, L. Manchanda, and M. T. Tang, Appl. Phys. Lett. **65**, 848 (1994).

[4] J. Ahn, J. Kim, G. Q. Lo, and D.-L. Kwong, Appl. Phys. Lett. **60**, 2089 (1992).

[5] H. Niimi, H. Y. Yang, and G. Lucovsky, in *Characterization and Metrology for ULSI Technology: 1998 International Conference*, edited by D. G. Seiler, A. C. Diebold, W. M. Bullis, T. J. Shaffner, R. McDonald, and E. J. Walters (The American Institute of Physics, Woodbury, NY, 1998), p. 273.

[6] Y. Ma, T. Yasuda, and G. Lucovsky, J. Vac. Sci. Technol. A **11**, 952 (1993); Y. Ma, T. Yasuda, S. Habermehl, and G. Lucovsky, J. Vac. Sci. Technol. B **11**, 1533 (1993); Y. Ma and G. Lucovsky, *ibid.* **12**, 2504 (1994).

[7] S. V. Hattangady, H. Niimi, and G. Lucovsky, J. Vac. Sci. Technol. A **14**, 3017 (1996).

[8] S. V. Hattangady, H. Niimi, and G. Lucovsky, Appl. Phys. Lett. **66**, 3495 (1995).

[9] S. V. Hattangady, R. Kraft, D. T. Grider, M. A. Douglas, G. A. Brown, P. A. Tiner, J. W. Kuehne, P. E. Nicollian, and M. F. Pas, Tech. Dig. Int. Electron Devices Meet. 495 (1996).

[10] Y. Wu and G. Lucovsky, in *Characterization and Metrology for ULSI Technology: 1998 International Conference*, edited by D. G. Seiler, A. C. Diebold, W. M. Bullis, T. J. Shaffner, R. McDonald, and E. J. Walters (The American Institute of Physics, Woodbury, NY, 1998), p. 293.

[11] C. G. Parker, G. Lucovsky, and J. R. Hauser, IEEE Electron Device Lett. **19**, 106 (1998).

[12] Y. Wu and G. Lucovsky, IEEE Electron Device Lett. **19**, 367 (1998).

[13] S. C. Song, H. F. Luan, Y. Y. Chen, M. Gardner, J. Fulford, M. Allen, and D. L. Kwong, Tech. Dig. Int. Electron Devices Meet. 373 (1998); H-H- Tseng *et al.*, *ibid.* 793 (1998).

[14] T. P. Ma, IEEE Trans. Electron Devices **45**, 680 (1998).

[15] G. Lucovsky, D. V. Tsu, R. A. Rudder, and R. J. Markunas, in *Thin Film Processes II*, edited by J. L. Vossen and W. Kern (Academic, Boston, MA, 1991), pp. 565–620.

[16] T. Yasuda, Y. Ma, S. Habermehl, and G. Lucovsky, Appl. Phys. Lett. **60**, 434 (1992).

[17] G. Lucovsky, H. Niimi, K. Koh, D. R. Lee, and Z. Jing, in *The Physics and Chemistry of SiO_2 and the $Si-SiO_2$ Interface*, edited by H. Z. Massoud, E. H. Poindexter, and C. R. Helms (Electrochemical Society, Pennington, NJ, 1996), p. 441.

[18] Z. Lu, M. J. Williams, P. F. Santos-Filho, and G. Lucovsky, J. Vac. Sci. Technol. A **13**, 607 (1995).

[19] G. Lucovsky, P. Santos-Filho, G. Stevens, Z. Jing, and A. Banerjee, J. Non-Cryst. Solids **198-200**, 19 (1996).

[20] G. Lucovsky and J. C. Phillips, J. Non-Cryst. Solids **227**, 1221 (1998).

[21] Y. Ma, T. Yasuda, and G. Lucovsky, Appl. Phys. Lett. **64**, 2226 (1994).

[22] B. C. Smith and H. H. Lamb, J. Appl. Phys. **83**, 7635 (1998).

[23] J. W. Kiester, J. E. Rowe, J. J. Kododzie, H. Niimi, H.-S. Tao, T. E. Madey, and G. Lucovsky, J. Vac. Sci. Technol. A (these proceedings); H. Niimi and G. Lucovsky, J. Vac. Sci. Technol. B (submitted).

[24] V. Misra *et al.*, IEEE Trans. Electron. Devices (submitted).

[25] Y-D. Wu and G. Lucovsky, IEEE Electron Device Lett. (submitted).

[26] G. Lucovsky, Y. Wu, H. Niimi, V. Misra, and J. C. Phillips, Appl. Phys. Lett. **74**, 2005 (1999).

[27] J. C. Phillips, J. Non-Cryst. Solids **34**, 153 (1979); **47**, 203 (1983).

[28] E. M. Vogel, K. Z. Ahmed, B. Hornung, W. K. Henson, P. K. McLartry, G. Lucovsky, J. R. Hauser, and J. J. Wortman, IEEE Trans. Electron Devices **45**, 1350 (1998).

[29] H. Y. Yang, H. Niimi, and G. Lucovsky, J. Appl. Phys. **83**, 2327 (1998).

[30] Y. Wu and G. Lucovsky, IEEE Trans. Electron Devices (submitted).

[31] H. Niimi and G. Lucovsky, J. Vac. Sci. Technol. B (submitted).

[32] J. Robertson and C. W. Chen, Appl. Phys. Lett. **74**, 1168 (1999).

[33] H. Z. Henisch, *Semiconductor Contacts* (Clarendon, Oxford, 1984), Chap. 2.

[34] A. Chatterjee *et al.*, Tech. Dig. Int. Electron Devices Meet. 779 (1998).

[35]H. Matsuhashi *et al.*, Jpn. J. Appl. Phys., Part 1 **33**, 1293 (1994).

[36]A. Rose, *Concepts in Photoconductivity* (Interscience, New York, 1963), Chap. 8.

[37]S. A. Campbell, D. C. Gilmer, X.-C. Wang, M.-T. Hsieh, H.-S. Kim, W. L. Gladfelter, and J. Yan, IEEE Trans. Electron Devices **44**, 104 (1997).

[38]B. Claflin, M. Binger, and G. Lucovsky, J. Vac. Sci. Technol. A **16**, 1757 (1998).

[39]J. H. Stathis and D. J. DiMaria, Tech. Dig. Int. Electron Devices Meet. 167 (1998).

[40]E. M. Vogel, W. L. Hill, V. Misra, P. K. McLarty, J. J. Wortman, J. R. Hauser, P. Morfouli, G. Ghibaudo, and T. Ouisse, IEEE Trans. Electron Devices **43**, 753 (1996).

[41]F. L. Galeener, W. Stutius, and G. T. McKinley, in *The Physics of MOS Insulators*, edited by G. Lucovsky, S. T. Pantelides, and F. L. Galeener (Pergamon, New York, 1980), p. 77.

[42]G. Lucovsky, A. Rozaj-Brvar, and R. F. Davis, in *The Structure of Non-Crystalline Materials 1982*, edited by P. H. Gaskell, J. M. Parker, and E. A. Davis (Taylor and Francis, London, 1983), p. 193.