PERGAMON

Microelectronics Reliability 40 (2000) 557–562

MICROELECTRONICS RELIABILITY

www.elsevier.com/locate/microrel

The relentless march of the MOSFET gate oxide thickness to zero

G. Timp [a,*], J. Bude [a], F. Baumann [a], K.K. Bourdelle [a], T. Boone [a], J. Garno [a], A. Ghetti [a], M. Green [a], H. Gossmann [a], Y. Kim [a], R. Kleiman [a], A. Kornblit [a], F. Klemens [a], S. Moccio [a], D. Muller [a], J. Rosamilia [a], P. Silverman [a], T. Sorsch [a], W. Timp [b], D. Tennant [a], R. Tung [a], B. Weir [a]

[a] *Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974-0636, USA*

[b] *University of Illinois, Urbana, IL 61801, USA*

Abstract

The narrowest feature of an integrated circuit is the silicon dioxide gate dielectric (3–5 nm). The viability of future CMOS technology is contingent upon thinning the oxide further to improve drive performance, while maintaining reliability. Practical limitations due to direct tunneling through the gate oxide may preclude the use of silicon dioxide as the gate dielectric for thicknesses less than 1.3 nm, however. © 2000 Elsevier Science Ltd. All rights reserved.

1. Introduction

An integrated circuit (IC) is a conglomerate of a large number of practically identical, very reliable switches interconnected with wires to express a computing function. The quintessential IC design is implemented using complementary pairs of silicon n- and pMOSFETs as the switches. Over the last 30 years, the cost to manufacture a transistor along with the wires necessary to incorporate it into a circuit has decreased by a factor of 100,000. It is anticipated that this unprecedented increase in productivity will continue unabated for the next 15 years.

One of the primary means for reducing cost, improving performance and increasing the scale of integration in an IC has been miniaturization. The guiding principle for miniaturization has been the scaling of the classical transport equations that describe the operating characteristics of an existing device to smaller dimensions to predict the characteristics of the smaller device with minimal design investment. An accurate appraisal of the limiting performance obtained by scaling CMOS technology is necessary, not only to identify the main impediments, but also to spur the timely development of alternatives. Currently, the critical dimensions of a MOSFET are the gate length (180 nm), the p–n junction depth (100 nm), and the gate oxide thickness, t_{ox} (3–5 nm); the narrowest feature is the gate oxide.

Here, we identify the gate oxide as the principal impediment to scaling the gate length of nMOSFETs and pMOSFETs to 35 nm, as shown in Fig. 1.

The viability of sub-50 nm CMOS technology is contingent upon the drive current performance. Improvements in the drive performance can be used to derate the power supply voltage, thereby improving reliability, reducing power dissipation, and enabling ultra-large scale integrated circuit (ULSI). The drive performance of a MOSFET is dictated by the SiO_2 gate dielectric thickness, and by carrier scattering in the channel. In what follows, we show that gate leakage current due to direct tunneling through the gate oxide will render SiO_2 thicknesses less than 1.3 nm impractical [1]. Consequently, the drive performance for $t_{ox} > 1.3$ nm is essentially limited by ballistic transport in the channel.

2. Gate leakage current

The motivation for reducing t_{ox} is apparent in Fig. 2(a) and (c); the observed increase in capacitance

[*] Corresponding author. Tel.: +1-908-582-4622; fax: +1-908-582-6000.

E-mail address: glt1@physics.bell-labs.com (G. Timp).

0026-2714/00/$ - see front matter © 2000 Elsevier Science Ltd. All rights reserved.
PII: S0026-2714(99)00257-7

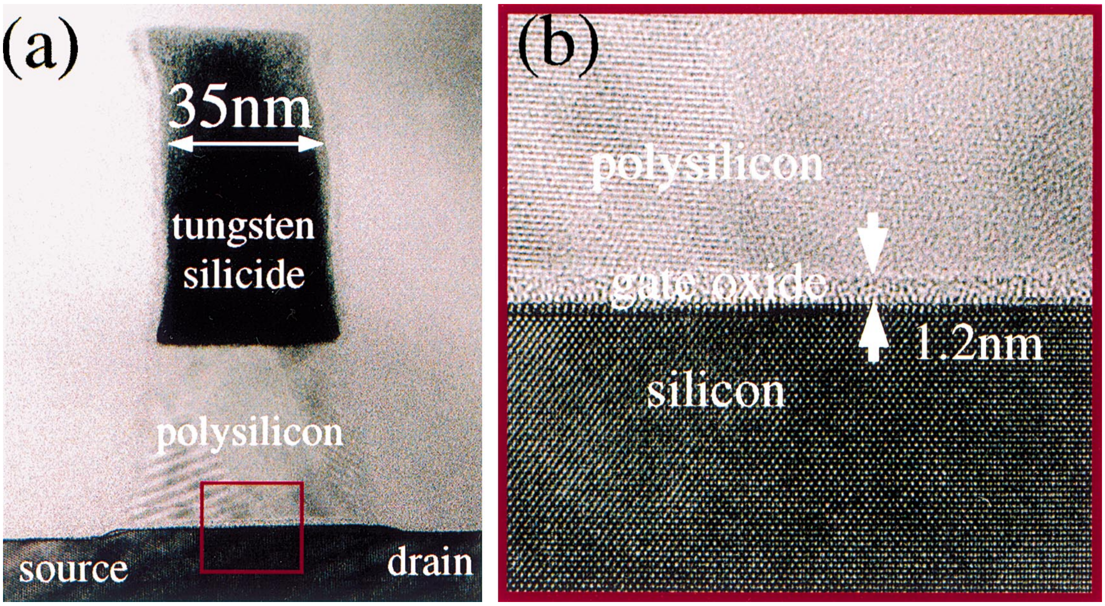

Fig. 1. (a) A high resolution transmission electron micrograph of the cross-section through a 35 nm gate length nMOSFET prior to side wall etch. The inset reveals a gate oxide thickness of approximately 1.2 ± 0.3 nm.

translates into increased drive current capability. Thus, ultra-thin gate oxides make a low voltage power supply practical. However, concomitant with the increase in the capacitance is a detrimental exponential increase in leakage current due to direct tunneling, as shown in Fig. 2(b) and (d). Juxtaposed with the measured exponential increase with voltage and t_{ox} observed in the MOS capacitors of Fig. 2(b) and (d) are the self-consistent quantum mechanical calculations of the tunneling current, which account for the measured doping profiles and the quantization of the accumulation layer. The agreement achieved over the range 1.3–3.1 nm is remarkable, considering that (1) the only free parameter is the t_{ox} (assuming a 3.15 eV electron barrier with a 0.5 m electron mass, and a 4.8 eV hole barrier with a 0.42 m hole mass), which nevertheless compares favorably to the measured ellipsometric value; and (2) that the trap density, $N_t = 0$.

A gate leakage current in excess of 1 A/cm^2 may preclude the use of oxides thinner than 1.3–1.4 nm, because the off-state power dissipation becomes comparable to the active power. The prospect of improving the tunneling characteristics by producing oxides with a lower trap density seems unlikely. The correspondence between the simulations and the measured leakage current indicate that the current at $V_g = -1.5$ V is dominated by the direct tunneling, not trap assisted tunneling [2]. As Fig. 3 indicates the data are consistent with a trap density cross-section product of 4 cm^{-1} over the range 1.1–3.1 nm, or a trap density of 4×10^{14} cm^{-3} for $\sigma = 1$ nm^2.

To explore the possibility of improving the tunneling characteristics by reducing interface roughness, we used electron energy loss spectroscopy in conjunction with scanning transmission electron microscopy (STEM) to map, with atomic resolution, the unoccupied electronic density of states (DOS) by site and atom column [3]. In this way, we can characterize the electrical and structural properties of the gate stack simultaneously. To improve contrast and sensitivity, we used the O–K edge to provide information about the unoccupied O-p contribution to the DOS. Fig. 6 shows the oxygen profile inferred from a 1.1 nm (ellipsometry) thick gate stack. An oxygen signal indicative of the bulk SiO_2 DOS is observed in a region approximately 0.8–1.0 nm wide, while a signal indicative of interfacial oxygen is located about either interface ranging in thickness from 0.3–0.5 nm wide.

At least two of the five silicon atoms, which span the oxide thickness are associated with the silicon/oxide interfaces, and the interfacial atoms have very different electrical and optical properties from the desired bulk silicon dioxide. In particular, the additional electronic states associated with the interfacial regions, which appear at energies below the bulk SiO_2 conduction band, are roughly aligned with the bulk Si conduction band and imply an altered dielectric response there (which may explain the discrepancies we

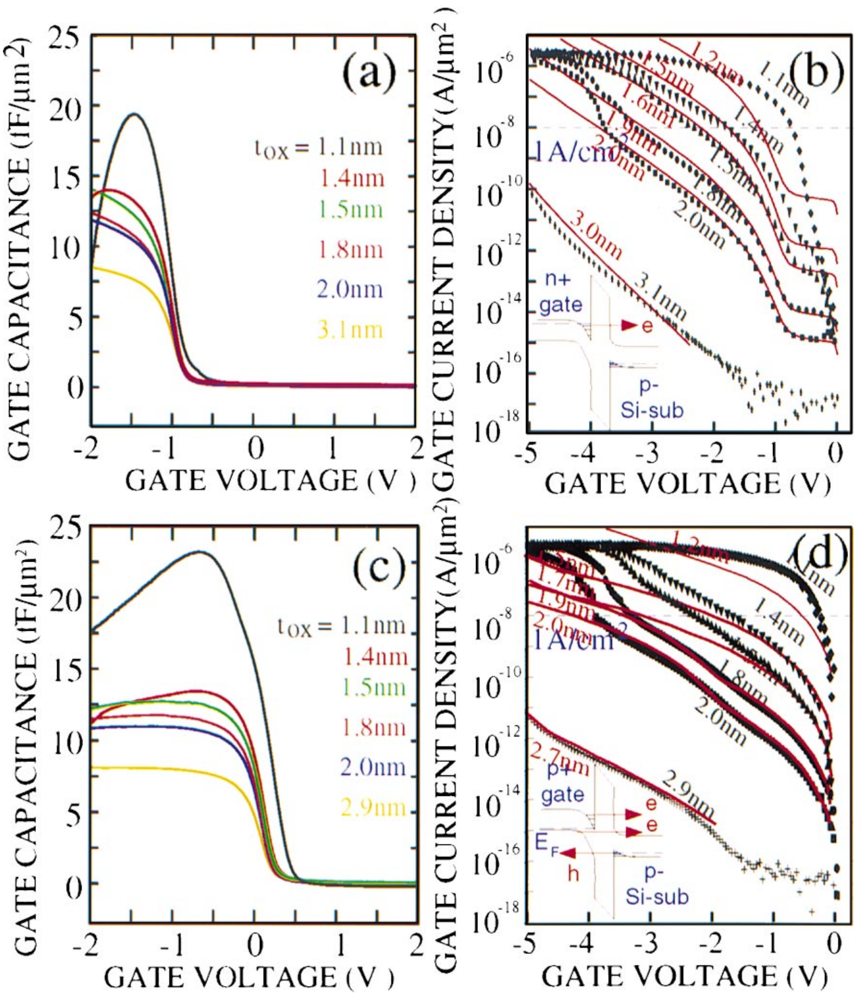

Fig. 2. (a) The capacitance versus gate voltage measured at 1 MHz on 200 μm × 200 μm capacitors with n+ poly gates on p-substrates with gate oxide thickness as a parameter (ellipsometry measurements shown as the black numbers). (b) The measured leakage current versus gate voltage (black symbols) and the corresponding simulations (red). The high frequency *C–V* measurements are compromised for $t_{ox} < 1.3$ nm, because the admittance is dominated by the conductance due to direct tunneling. The only free parameter in the quantum mechanical simulations is the oxide thickness (red numbers). (c) The capacitance versus gate voltage measured at 1 MHz on 200 μm × 200 μm capacitors with p+ poly gates on p-substrates with oxide thickness as a parameter (ellipsometry measurements indicated by the black numbers). (d) The leakage current versus gate voltage (black symbols) on the same substrate. The corresponding simulations of the leakage current are shown in red. These simulations account, not only for the measured doping distribution and quantization of the accumulation layer, but also treat the hole tunneling and electron tunneling from the conduction and valence band.

observe between t_{ox} measured using ellipsometry versus TEM.) These gap states result from the exponential decay of the silicon conduction band wavefunction into the oxide, and the tunneling current depends on the overlap of these states. An adequate tunneling barrier must be at least 6λ, where λ is the decay length of the evanescent state. The minimum thickness for an ideal oxide barrier is about 0.7 nm. Interfacial roughness contributes at least another 0.6 nm, which puts a lower limit of 1.3 nm on a practical gate oxide thickness.

3. Boron penetration

The most stringent requirements for high performance sub-50 nm CMOS technology are dictated by the PMOS transistor. In addition to satisfying the gate

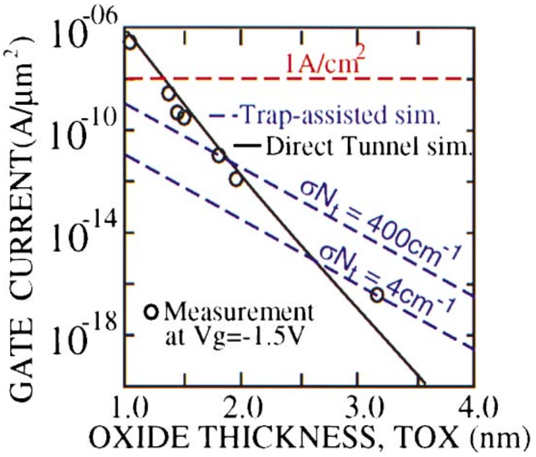

Fig. 3. Gate leakage current plotted as a function of oxide thickness. The open circles represent measurements taken at $V_g = 1.5$ V using n+ poly on p-type substrate capacitors. The solid (black) line represents the direct tunneling simulation; the (blue) dashed lines represent trap-assisted tunneling simulations.

leakage criterion, the pMOSFET gate oxide must resist boron penetration from the heavily doped polysilicon gate electrode through the oxide into the underlying channel to avoid threshold variations. The absence of a shift in the flat-band voltage with t_{ox} in Fig. 2(c) and the linear dependence of the threshold voltage t_{ox} shown in Fig. 4 indicate minimal boron penetration through oxides as thin as 1.0 nm. This resilience of the gate oxide to boron penetration is consistent with other estimates of the boron diffusivity ($D_b < 5 \times 10^{-17}$ cm^2/s) and segregation coefficient [4]. To corroborate these results, a SIMS analysis was performed to investigate boron penetration in oxides ranging from 1.0–10 nm thickness. The SIMS data show no indication of excessive boron penetration into the underlying substrate for oxides as thin as 1.0 nm relative to the 10 nm gate oxide, following a thermal cycle of 1050°C for 5 s (Fig 5). Boron penetration is observed through a 1.0 nm thick oxide for a 1050°C 10 s cycle, however.

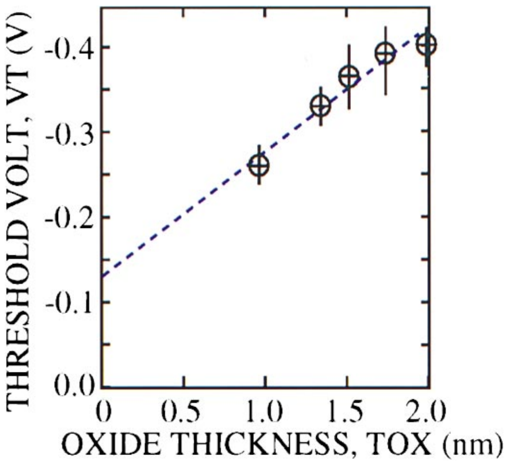

Fig. 4. The threshold voltage of 165 ± 10 nm gate pMOSFETs versus oxide thickness. A linear dependence extrapolates to $t_{ox} = 0.9$ nm and indicates minimal poly-depeletion and boron penetration following a 1000°C, 5 s anneal.

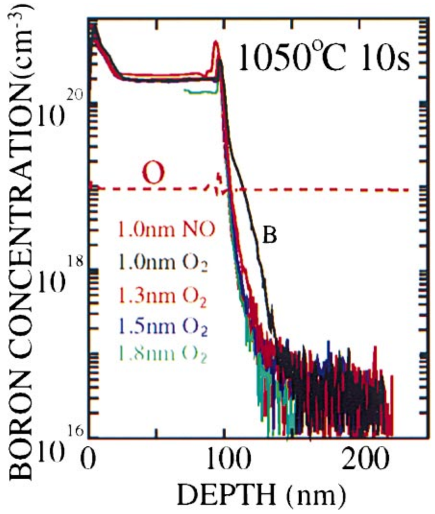

Fig. 5. SIMS analysis showing the boron profiles through a cross-section of a capacitor following an anneal at 1050°C for 10 s. The profile corresponding to the 1.0 nm thick SiO_2 layer indicates boron penetration, however the 1.0 nm thick SiO_xN_y layer shown minimal penetration for this thermal budget.

4. The ballistic nanotransistor

If the gate leakage current renders SiO_2 thicknesses <1.3 nm impractical, then the drive performance of a MOSFET is limited by ballistic transport in the channel. The correspondence between our measurements of the DC drain characteristics of sub-50 nm gate-length nMOSFETs and 3D full band Monte-Carlo simulations show that ballistic transport can be achieved for effective channel lengths <30 nm. To characterize the transport, we use as a metric the transmission probability (T) of a net flux emanating from the source to reach the drain. In the degenerate limit, the saturated drain current is given by $I_{Dsat} = T/(2-T)qn\langle v(n)\rangle$, where $\langle v(n)\rangle$ is the average thermal carrier velocity for a degenerate gas with density n, where $n = C_{eff}(V_g - V_t)$ and C_{eff} is the effective gate capacitance. For transverse fields >0.5 MV/cm, the scattering in the channel is dominated by interface

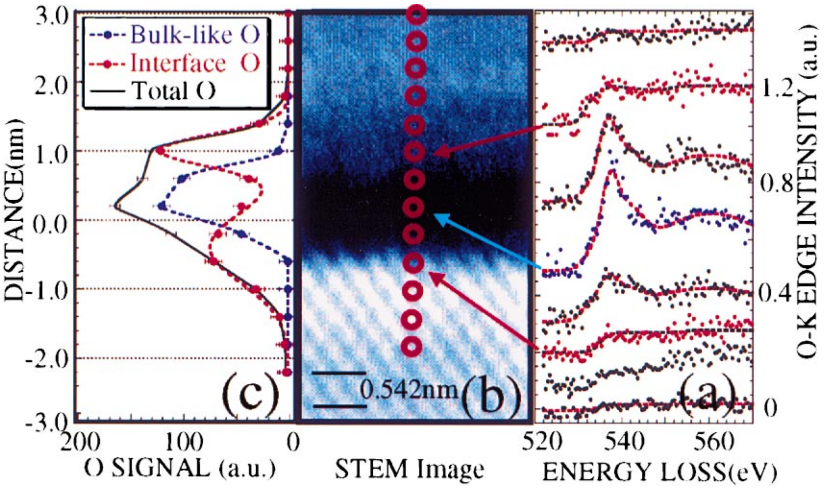

Fig. 6. (a) The EELS O–K edge measured across a gate stack with a nominally 1.1 nm thick gate oxide. The annular dark field image in (b) shows where each of the spectra were taken. The silicon substrate is at the bottom, the oxide is in the middle, and a deposited a-Si layer is in the top half. The smooth curves in (c) are the best fits using a mixture of the bulk and interfacial spectra. The nominally 1.1 nm thick gate oxide shows an oxygen signal distributed across 1.6 nm with a bulk SiO_2 signal found only in a 0.8–1.0 nm thick region near the center. The interface roughness is <0.5 nm.

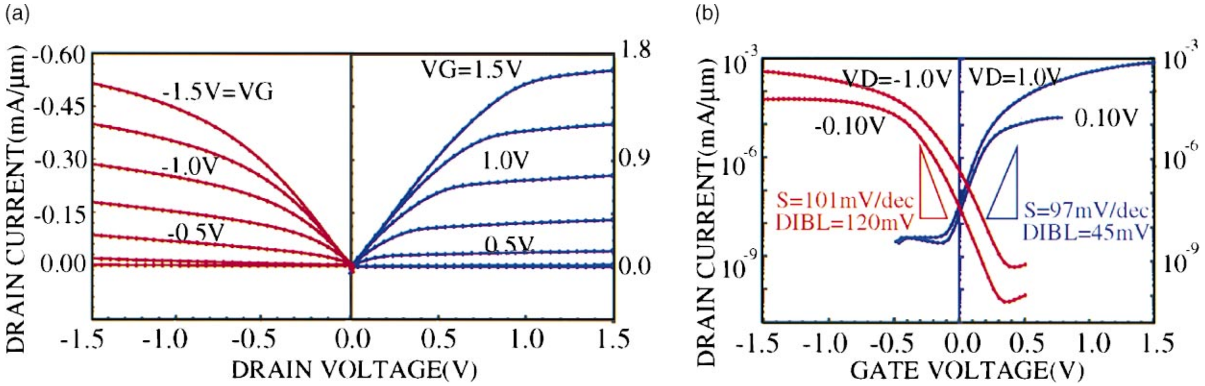

Fig. 7. The (a) drain and (b) subthreshold characteristics of a 40 nm gate length nMOSFET (blue) with a 1.6 nm gate oxide, and a 40 nm pMOSFET (red) with a 1.5 nm gate oxide. The nMOSFET has a drive current of $I_{\mathrm{Dsatn}} = 0.73$ mA/μm at $V_D = V_G = 1$ V with a threshold voltage of $V_{\mathrm{tn}} = 0.36$ and a gate leakage < 5 nA/μm² at $V_G = 1$ V. For the pMOSFET, $I_{\mathrm{Dsatp}} = 0.28$ mA/μm at $V_D = V_G = -1$ V with $V_{\mathrm{tp}} = -0.29$ V and gate leakage < 2 nA/μm².

roughness like that illustrated in Fig. 6. The drain and subthreshold characteristics measured in sub-50 nm nMOSFET and pMOSFET with a 1.5–1.6 nm thick gate oxide and a 1.5 V power supply are shown in Fig. 7. The transmission, corresponding to the drain current at 1.5 V is $0.86 < T < 0.95$ for the nMOSFET and <0.5 for the pMOSFET indicating that a 20%/300% improvement in the nMOSFET/pMOSFET I_{Dsat} may still be possible, provided $T = 1$ can be achieved. The fundamental nature of the interface roughness, illustrated in Fig. 6, demands that the transverse field be reduced to promote higher transmission.

References

[1] Timp G, Bourdelle KK, Bower JE, Baumann FH, Boone T, Cirelli R, Evans-Lutterodt K, Garno J, Ghetti A, Gossmann H, Green M, Jacobson D, Kim Y, Kleiman R, Klemens F,

Kornblit A, Lochstampfor C, Mansfield W, Moccio M, Muller DA, Ocola L, O'Malley M, Rosamilia J, Sapjeta J, Silverman P, Sorsch T, Tennant DM, Timp W, Weir BM. IEDM Tech Dig, 1998. p. 615–8.

[2] Ghetti A, Hamad A, Silverman PJ, Vaidya H, Zhao N. Proceedings of the 1999 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD). p. 239.

[3] Muller DA, Sorsch T, Moccio S, Baumann FH, Evans-Lutterodt K, Timp G. Nature 1999;399:758–61.

[4] Aoyama T, Suzuki K, Tashiro H, Tada Y, Arimoto H. IEDM Tech Dig, 1997. p. 627–30.