Invited Paper

Gate Oxide Scaling Limits and Projection

Chenming Hu
Dept. of Electrical Engineering and Computer Sciences
University of California, Berkeley, CA 94720

Abstract

MOSFET gate oxide scaling limits are examined with respect to time-dependent breakdown, defects, plasma process damage, mobility degradation, poly-gate depletion, inversion layer thickness, tunneling leakage, charge trapping, and gate delay. It is projected that the operating field will stay around 5MV/cm for reliability and optimum speed. Tunneling leakage prevents scaling below 2nm, which is sufficient for MOSFET scaling to 0.05µm.

Introduction

In order for a MOSFET to behave as a transistor, the gate must exert greater control over the channel than the drain dose, i.e., the gate to channel capacitance must be larger than the drain to channel capacitance. A simple model suggests (1)

$$L_{min} \propto T_{ox} \bullet X_j^{1/3} \qquad [1]$$

A survey of the literature would reveal that the most thorough device physics studies of the late 1970's pegged the scaling limit of MOSFET at 0.5µm. In the mid 1980's, it was 0.25µm. Today, no device physics barrier is foreseen for scaling to at least 0.1µm. What fogged the crystal balls of the yesteryears was the uncertainty over the minimum acceptable T_{ox} in Eq. [1]. The scaling limit of T_{ox} is therefore of paramount importance. Besides suppressing the short channel effect, reducing T_{ox} improves I_d and generally but not always raises circuit speed. Thinner tunnel oxide would also be desirable for lowering the program voltage of nonvolatile memory. Clearly, there are many strong incentives to reduce T_{ox} at each technology generation. What, then, are the limits to oxide scaling? This paper attempts to answer this critical question.

Oxide Breakdown

Oxide breakdown has historically been the limiting factor in choosing T_{ox}. The past pessimistic predictions of L_{min} can be directly attributed to a lack of understanding of the oxide breakdown limit.

•*Intrinsic Breakdown Field.* If gross "defects" are not present, i.e., if one studies oxide samples which are very much smaller than 1mm^2 in size, the lifetime of the oxide is surprisingly predictable (2). Temperature effect (3), a physical interpretation and refinement for very thin oxide(4) have been presented. The model predicts the oxide lifetime as a function of T_{ox} and V_{ox} very well (Fig. 1) (4).

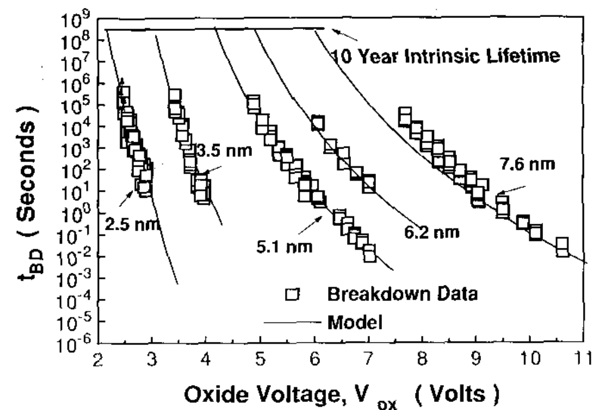

Fig. 1 Oxide lifetime has been described by a hole-injection model.

Both data extrapolation and the model predict that oxide can have 20 years lifetime at 125°C up to oxide field of 7MV/cm, 8MV/cm for below 5V operation. In Fig. 2, for example, 5.5V operation can use 8nm, 3.6V operation requires only 4.5nm, and 2.75V requires 3.3nm.

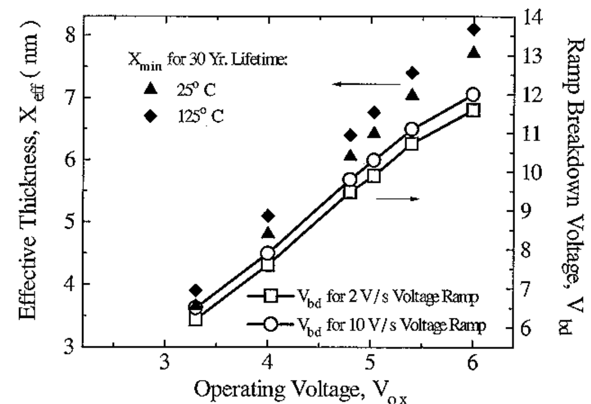

Fig. 2 The maximum acceptable field for 30 year lifetime of defect-free oxide is 7-8 MV/cm.

0-7803-3393-4/96/$5.00 (c) 1996 IEEE

•*Derating of Breakdown Field due to Defects.* A manufacturable process must provide a soft margin above the intrinsic breakdown limit. Recent production experiences suggest that ~30% derating to 5 to 5.5MV/cm seems to be reasonable, i.e., 11nm for 5.5V, 6.5nm for 3.6V, and 4.5nm for 2.75V.

For a given manufacturing line, one can use an "effective thickness" defect model to predict the product oxide yield, reliability failure rate, optimal burn-in condition, etc. from the statistical distribution of the breakdown voltage of oxide test samples (5).

•*Process Induced Damage.* Fig. 3 shows that 18nm oxide with large "antenna" structure has low interface trap density after plasma etch indicating little plasma-induced damage. 11.6nm oxide has an order of magnitude more interface traps, reflecting significant damage. Surprisingly, the very thin 6.4nm oxide exhibits much less damage than the 11.6nm oxide.

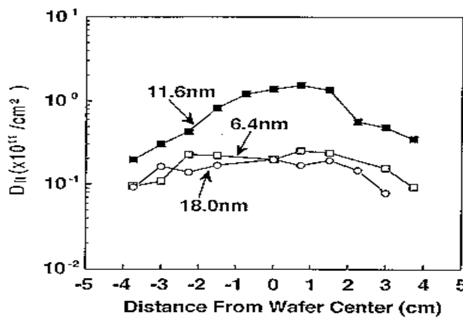

Fig. 3 Plasma process antenna-effect actually causes less damage to 6.4 nm oxide than 11.6 nm oxide.

This exciting trend had first been predicted by a model illustrated in Fig. 4 (6). Langmuir equation describes the plasma current (I_p) collected by the antenna (plasma probe).

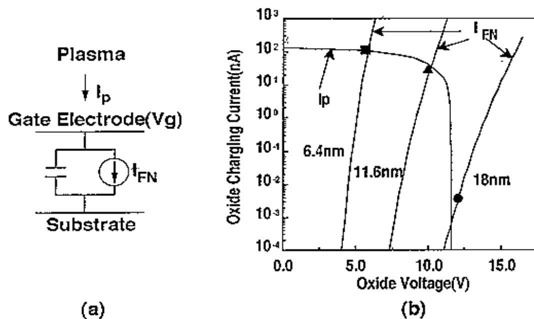

Fig. 4 Plasma charging (I_P) resembles a constant-current source rather than a fixed voltage source, therefore does not devastate very thin oxides. Included are three data points.

Fowler-Nordheim equation describes the oxide current, I_{FN}. The intersections determine the stress currents and voltages in this plasma process equipment. Clearly, the oxide stress current hardly increases as T_{ox} is reduced from 10 nm to 0 nm. To understand Fig. 3, one needs to also realize that thinner oxide (6.4nm) is more tolerant of the same current stress than thicker (11.6nm) oxide.

Transistor Current and Speed

All else being equal, MOSFET current always increases when T_{ox} is reduced, although at a lower rate than a simple model might suggest because of mobility reduction, polysilicon gate depletion, and finite inversion layer thickness. Gate speed, on the other hand, may slow down due to excessive T_{ox} reduction.

•*Mobility.* Electron and hole mobilities have been shown to be highly predictable function of $(V_g+V_t)/T_{ox}$ as shown in Fig. 5 (7). Once $V_g(=V_{dd})$, V_t, and T_{ox} are specified, mobilities are known. See (7) for modification for buried channel MOSFET.

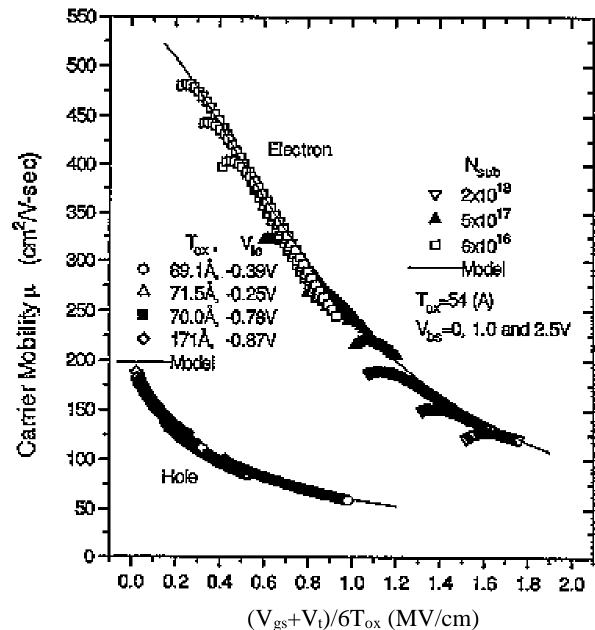

Fig. 5 Carrier mobility remains unchanged, if $V_{dd,}$ $V_{t,}$ and T_{ox} are scaled in proportion.

•*Polysilicon Gate Depletion and Inversion Layer Quantization.* Gate depletion can be simply modeled with the usual MOS (polysilicon gate being S) depletion equation (8). The big surprise, still unexplained, is that the polysilicon doping concentration at the oxide interface seems to be only around $3x10^{19}$ cm^{-3}. At E_{ox}=5MV/cm, the depletion layer thickness is about 0.6nm equivalent oxide thickness or 0.3V reduction in Vg. This is a significant

"excess oxide thickness" as we reduce Tox to 6nm or below. Even with the polysilicon gate biased into accumulation, the CV data would indicate an "electrical T_{ox}" that is about 0.5nm thicker than the physical T_{ox} as determined by optical techniques or from tunneling IV characteristics. This is because the inversion layer carriers are located at about 1.5nm (0.5nm effective oxide thickness) below the Si/SiO_2 interface. The middle curve in Fig. 6 (9) is the I_{dsat} predicted by today's standard 2D device simulators for some future technologies. The bottom curve is obtained when polygate depletion and inversion layer quantization are included in the simulation. As an important beside, the top curve includes polygate depletion, inversion layer quantization, and velocity overshoot (energy balance) effect (9).

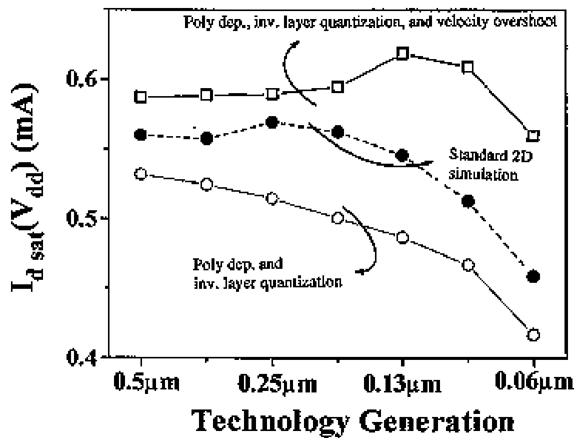

Fig. 6 Middle and bottom curves show the effect of polysilicon depletion and large inversion layer thickness due to quantization, Some reasonable V_{dd} and T_{ox} values are used.

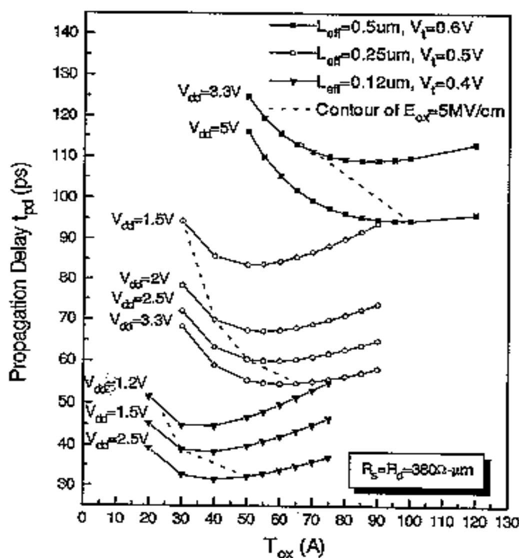

Fig. 7 At low V_{dd}, e.g. 2 Volts and 0.25 μm, optimum speed may require a thicker T_{ox} than that allowed by the 5 MV/cm limit.

•*Speed.* Fig. 7 shows the projected inverter delay time versus the T_{ox} (7). For low V_{dd} operation, maximum speed may dictate the use of an oxide that is thicker than what the 5MV/cm breakdown limit would allow. While the optimal thickness should be thinner than what Fig. 7 suggests if the circuit is more heavily loaded by the interconnect capacitance, anecdotal data of large CPUs actually agree with Fig. 7 well.

Oxide Leakage and Device Drift

•*Direct Tunneling.* Whenever oxide voltage is lower than 3.2V (the Si/SiO_2 barrier voltage), the electron tunneling barrier changes from being triangular to trapezoidal and the oxide current, known as the direct tunneling current, remains high at even 1V and is very sensitive to T_{ox} as shown in Fig. 8 (4). Static logic circuits can tolerate large gate leakage, e.g. $1A/cm^2$ (even though the junction leakage is typically $1\mu A/cm^2$). For 1V operation, Fig. 8 would suggest 2 or 2.5nm as the scaling limit. DRAM can tolerate less oxide leakage and typically bootstraps above V_{dd}. 3nm may be the T_{ox} limit for DRAM, making scaling below 0.1μm difficult. This discrepancy needs a resolution.

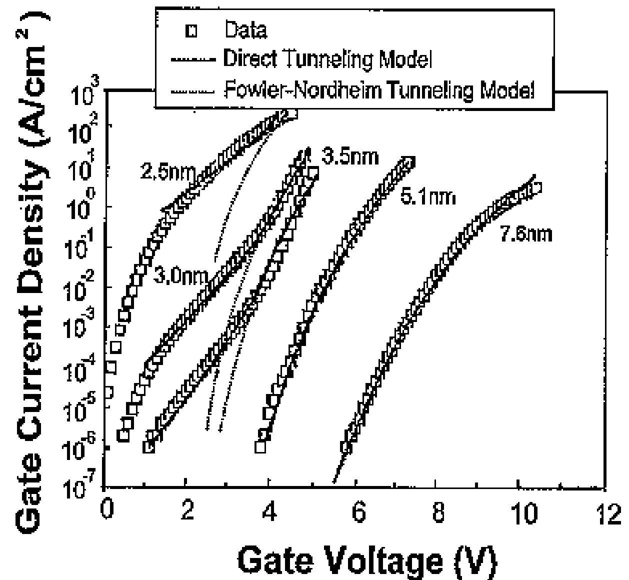

Fig. 8 Direct tunneling may limit oxide scaling to around 2.5 nm.

•*Stress Induced Leakage.* High field stress of thin oxide creates low-field leakage, apparently through the generation of neutral oxide traps that facilitate electron tunneling. Fig. 9 highlights the different behaviors of thick and thin oxides (10). This leakage makes nonvolatile memory tunneling oxide scaling much below 8nm difficult and perhaps impossible unless the 10 year charge retention requirement is relaxed (11).

0-7803-3393-4/96/$5.00 (c) 1996 IEEE

•*Charge Trapping and MOSFET Stability*. Even if the static logic circuits can tolerate very large oxide leakage and therefore very thin oxide, will there be severe charge trapping and MOSFET drift as a result of the huge charge

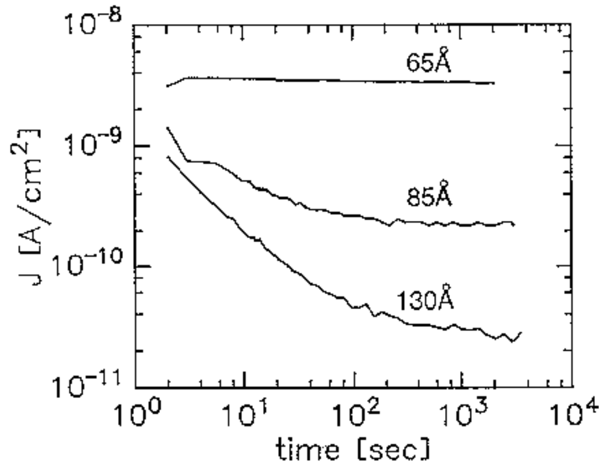

Fig. 9 Stress-induced leakage is a transient current in thicker oxides but a continuous current in thinner oxides, preventing NVM oxide scaling.

fluence in 20 year's time? Only preliminary studies have been reported (12) and early indication is that 3nm oxide can tolerate at least 1000 coul/cm^2, i.e. 20 years at 1V, of charge passage without significant drift.

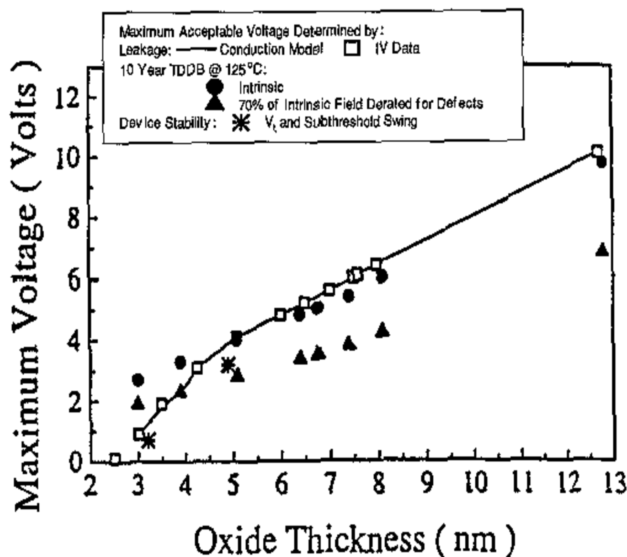

Fig. 10 Scaling limits due to breakdown, tunneling leakage, and device stability (12).

Summary and Projection

Fig. 10 and Fig. 7 summarize the consideration for choosing T_{ox}. 20 year intrinsic lifetime can be achieved at 7MV/cm in 8nm oxide, rising gradually to 9MV/cm in 3nm oxide. 30% derating for defects puts the field limit at 5MV/cm rising to 6 MV/cm. Below 2.5V, oxide leakage (Fig. 10) and circuit speed (Fig. 7), rather than oxide breakdown, will dictate the choice of the oxide thickness. (Oxide defects will still need to be vigilantly controlled.)

The optimal T_{ox} are:

~12nm if 5V/0.5µm, limited by defect breakdown;

~6.5nm if 3.3V/0.35µm, limited by defect breakdown and circuit speed;

~4.5nm if 2.5V/0.25µm, limited by defect breakdown and circuit speed ;

~4nm if 1.5V/0.18µm, limited by circuit speed;

~3nm if 1.5V/0.1µm, limited by circuit speed and defect breakdown;

~2nm if 1V/0.05µm, limited by leakage.

The above T_{ox} values are the final physical thickness. Electrical thickness would be 0.5nm thicker . Tox can be up to 30% thinner than those listed above if only small area of oxide is used (so that the margin allowance for defects can be small), e.g. 3.3V I/O devices in 2.5V products, etc. (13). Excess oxide leakage will limit scaling just above 2nm, which makes 0.05µm MOSFET possible. DRAM gate oxide thickness will reach limit earlier at ~3nm. Plasma process damage will not become much more serious with future oxide scaling.

Acknowledgment

This work is supported by SRC-IJ148, AFOSR, JSEP, AMD, TI, IDT, and MICRO.

References

(1) Z.H. Liu, et al, IEEE Trans. Electron Dev., p.86, 1993
(2) I.C. Chen, et al, IEEE Trans. Electron Dev., p.333, 1985
(3) R. Moazzami, et al, IEEE Trans. Electron Dev., p.2462, 1989
(4) K.F. Schuegraf, et al, Semiconductor Sceince and Technology, p.989, 1994
(5) R. Moazzami, et al, IEEE Trans. Electron Dev., p.1643, 1990
(6) H. Shin, et al, IEEE Electron Device Letters, p.509, 1993
(7) K. Chen, et al, IEEE Electron Device Letters, p.202, 1996
(8) K.F. Schuegraf, et al, Int'l Symp. on VLSI Technology, Systems and Appl., Taipei, p.86, 1993
(9) D. Sinitsky, et al, submitted to IEEE Electron Device Letters
(10) R, Moazzami, et al, IEDM, p.139, 1992
(11) C.H. Wann, et al, IEDM, p.867, 1995
(12) K. Schuegraf, et al, IEDM, p.609, 1994
(13) C. Hu, et al, ISSCC, p.86, 1994

0-7803-3393-4/96/$5.00 (c) 1996 IEEE