Extended Abstracts of the 1997 International Conference on Solid State Devices and Materials, Hamamatsu, 1997, pp. 6-7

Invited

Reliability and Scaling of Thin Gate Oxide

Chenming Hu

Dept. of Electrical Engineering & Computer Sciences, University of California, Berkeley, CA 94720, USA
Phone/Fax: 510-642-3393/510-642-2739, Email: hu@eecs.berkeley.edu

1. Introduction

In order for a MOSFET to behave as a transistor. the gate must exert far greater control over the channel than the drain dose, i.e., the gate to channel capacitance must be much larger than the drain to channel capacitance. In this sense. the channel length scaling limit will be largely determined by the limits of gate oxide thickness reduction.

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.

2. Oxide Breakdown

Oxide breakdown has historically been the limiting factor in choosing T_{ox}. The pessimistic predictions made in the past decades of the scaling limit of MOSFET channel length can be directly attributed to a lack of understanding of the oxide breakdown limit.

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 and the breakdown field of the oxide is surprisingly predictable [1] and quite insensitive to preoxidation surface treatment and the oxidation condition. This is the "intrinsic breakdown" of gate oxide. Temperature effect [2], a physical interpretation and refinement for very thin oxide [3] have been presented. The model predicts the oxide lifetime as a function of T_{ox} and V_{ox} very well (Fig. 1).

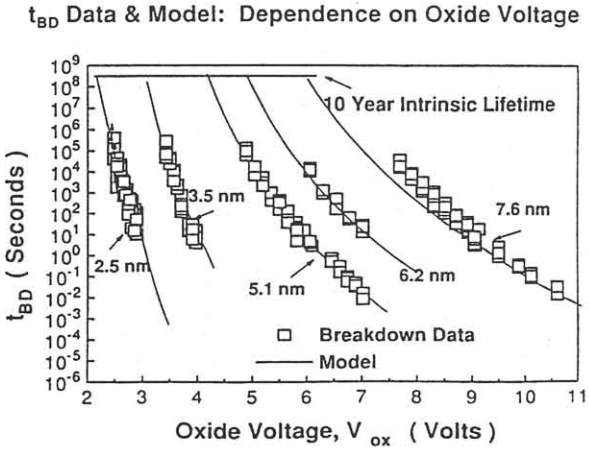

Fig. 1: Oxide lifetime has been described by a hole-injection model.

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 7.5nm, 3.6V operation requires only 4.5nm, and 2.75 V requires 3nm of intrinsic or gross-defect free gate oxides.

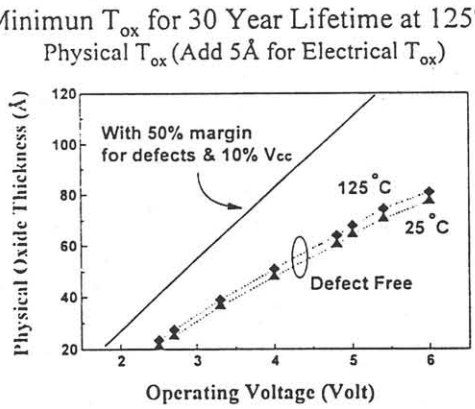

Fig. 2: The minimum acceptable oxide thickness for 30 year lifetime of defect-free (intrinsic) oxide and manufacturable ULSI gate oxide.

A volume manufacturing process must provide a certain margin of safety above the intrinsic oxide thickness limit on account of oxide defects. 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 [4]. Recent production experience suggests that it is adequate to use oxides 50% thicker than the intrinsic limits. That plus the 10% V_{cc} margin results in the solid line in Fig. 2 as a projection of manufacturable oxide thickness. The results are quite bright; 4nm is adequate for 2.5V. Very likely the oxide thickness choice will be influenced by considerations beyond oxide breakdown reliability.

3. Process Induced Damage

Fig. 3 shows that very thin gate oxides are not stressed as hard as the thicker oxides by plasma process [6]. This counterintuitive and welcome phenomenon was predicted by a model that treats the antenna as a Langmuir probe concluding that the plasma charging process resembles a fixed current source rather than a voltage source for very thin oxides [7]. In addition, thinner oxides can tolerate larger stress cur-

rents and charge densities. Minimizing the defect density in very thin oxide will remain a challenge.

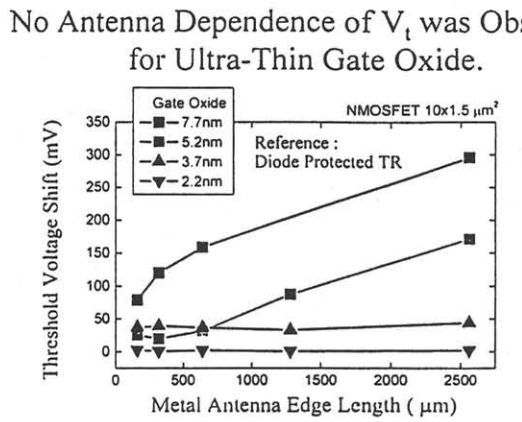

Fig. 3: Plasma process induced damage is actually less in very thin intrinsic oxides than in thicker oxides.

4. 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 [5]. As a result, there is an optimum range of oxide thickness for circuit speed performance [8].

5. 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} [9]. Static logic circuits can tolerate large gate leakage, e.g. $1A/cm^2$ (even though the junction leakage is typically 1 $\mu A/cm^2$). DRAM can tolerate less oxide leakage and typically bootstraps above V_{dd}. 3nm may be the T_{ox} limit for DRAM, making scaling of DRAM transistors more difficult.

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 [10]. This leakage makes non-volatile memory tunneling oxide scaling much below 7nm difficult and perhaps impossible unless the 10 year charge retention requirement is relaxed [11].

Charge Trapping and MOSFET Stability

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.

6. Summary

Gate oxide scaling will be determined by several factors summarized in Fig. 4. Assuming continued effort and success in manufacturing low defect-density oxide, oxide breakdown reliability may not be the limiting factor. Below 2.5V, circuit speed optimization will dictate a thickness larger than that necessary for acceptable breakdown reliability. Below 1V, direct tunneling will limit oxide scaling to 2nm. This is sufficient for 0.05 μm MOSFET and perhaps beyond.

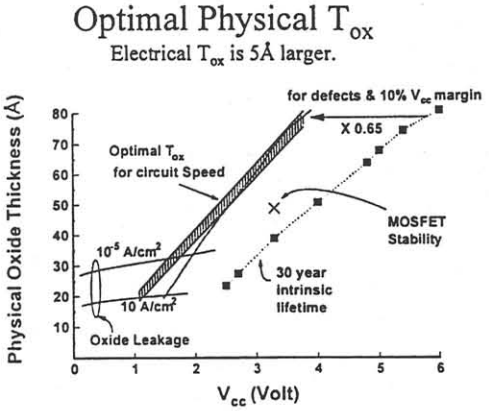

Fig. 4: Summary of gate oxide scaling considerations.

Acknowledgment

This work is supported by SRC-IJ148, AFOSR, ONR, NSF, TI, IDT, and MICRO.

References

1) I.C. Chen, et al, IEEE Trans. Electron Dev. (1985) p.333.
2) R. Moazzami, et al, IEEE Trans. Electron Dev. (1989) p.2462.
3) K.F. Schuegraf, et al, Semiconductor Science and Technology (1994) p.989.
4) R. Moazzami, et al, IEEE Trans. Electron Dev. (1990) p.1643.
5) C. Hu, IEDM (1996) p.319.
6) D.G. Park, et al, *Int'l Symp. on Plasma Process-Induced Damage* (June 1997) p.15.
7) H. Shin, et al, IEEE Electron Device Letters (1993) p.509.
8) K. Chen, et al, IEEE Electron Device Letters (1996) p.202.
9) K.F. Schuegraf, et al, *Int'l Symp. on VLSI Technology, Systems and Appl., Taipei* (1993) p.86.
10) R. Moazzami, et al, IEDM (1992) p.139.
11) C.H. Wann, et al, IEDM (1995) p.867.
12) K.F. Schuegraf, et al, IEDM (1994) p.609.