MOSFET DEVICE SCALING: A (BIASED) HISTORY OF GATE STACKS

C.M. Osburn and H.R. Huff*

North Carolina State University, Raleigh, NC 27695
*International SEMATECH, Austin, TX 78741

INTRODUCTION

Device scaling has been the engine driving the integrated circuit (IC) microelectronics revolution as described by Moore's Law [1]. The critical elements in device scaling are the gate dielectric thickness, the channel length, L_g, and the junction (and now) extension junction depth [2]. These dimensions have changed from their early 1970's values of 50-100 nm, 7.5 μm, and ~ 1 μm, respectively, to 1.1-1.6 nm, 45 nm and 25 nm (extension junction depth) for the high performance microprocessor (MPU) for the 100 nm Technology Generation as described in the International Technology Roadmap for Semiconductors (ITRS) [3]. The gate dielectric may indeed be **the** key structural element in the IGFET. It is the smallest dimensional element in a transistor and yet has to withstand the highest electric field. It must have low levels of fixed charge (~ 5 x 10^{10}/cm^2) and interface states (~ 5 x 10^{10}/cm^2-eV) and must remain reliable after years of high field stressing (~ 10 years) Since it occupies a large fraction of the total chip area and thus can dominate yield, it must also exhibit a very low defect density. Finally, it must be thermally, chemically, and mechanically compatible with the other materials and the manufacturing processes used in IC fabrication. Fortunately, silicon dioxide (SiO_2) possesses those attributes to the extent that no other material does. Although SiO_2 has been the mainstay of the industry for the past 40 years, remarkable technologies were needed to enable its use in the first place and considerable material advances were required after its introduction to allow its continued use, as the thickness has scaled from about 100 nm to 1 nm.

We briefly describe some of the early problems that needed to be solved to allow the use of SiO_2 and its implementation for device scaling. Improved SiO_2 fabrication and subsequent processing techniques were necessary to reduce the oxide thickness while preserving high yield and reliability. After serving so well for four decades, however, the limit of SiO_2 appears to be in sight. State-of-the-art oxides today (often mixed with low to moderate concentrations of nitrogen) are about 1.5 nm thick (corresponding to about five molecular layers thick, which includes two "SiO_x" layers for bonding to the silicon and polysilicon gate electrode). While it may be feasible to make SiO_2 thinner, direct tunnelling compromises their ability to remain an insulator. An intensive global search is now in progress to find an alternate gate dielectric as well as for the gate electrode.

THE EARLY YEARS (1955-1975)

The initial use of SiO_2 in practical semiconductor devices was for bipolar devices. The discovery by Frosch and Derrick in 1957 that a thin layer of SiO_2 can effectively mask the diffusion of most important dopants [4] led to its use for junction formation. The observation of junction passivation by Hoerni in a planar process in 1960 led to its use for improved junction characteristics [5]. The experience gained by fabrication of junctions via SiO_2 masking in bipolar devices led D. Kahng and M.M. Atalla to report the first MOSFET on a silicon substrate in 1960 [6]. Following J. Kilby's invention in 1958 of the IC using germanium [7], R. Noyce [8] quickly followed in 1959 utilizing SiO_2 grown on silicon. The SiO_2 was an extremely good insulator on which Al wiring could subsequently be adherently deposited.

The initial set of challenges associated with the use of SiO_2 included: control of the growth process (oxide thickness) and the reduction of mobile charges, fixed charge (Q_f) and interface state charge (D_{it}). While most of the industry initially selected to fabricate PMOS devices, some companies elected to fabricate NMOS because of the higher channel mobility for electrons. However, positive charge control is a more serious issue in NMOS technology—especially in the parasitic field oxide devices. Post-oxidation and post-metallization anneals were developed in the 1960's to minimize both fixed and interface charge [9,10]. The mobile charge, such as Na and K, required stringent control. Phosphosilicate glass was found to be an effective getter for mobile ions [11]; the later use of HCl in the oxidizing ambient was pioneered in the early '70s as a way to transport lifetime-killing metallic impurities and mobile ions from the wafer to the gaseous ambient [12, 13].

Once the initial issues in growing oxides and controlling charge were "solved," attention was directed to manufacturing and reliability issues, where dielectric breakdown was determined to be both a yield and a reliability limiter. Significant improvements in dielectric breakdown voltages were achieved by adding halogens such as HCl, to the oxidizing atmosphere [14]. Reliability improvements were observed when an appropriate amount of hydrogen was present in the dielectric (or at the silicon/SiO_2 interface). Nevertheless, reliability considerations limited the minimum dielectric thickness that was considered acceptable.

DEVICE SCALING (1975-2000)

The 1970's brought on the widespread use of dimensional scaling as the predominant means to increase device density and lower transistor cost via R. Dennard's scaling methodology [15]. Initially, dielectric thickness and operating voltage were to be scaled at the same rate so that the electric field across the gate dielectric remained constant; nevertheless, the industry soon adopted a constant voltage mode of scaling. The higher electric field associated with constant voltage scaling degraded the oxide integrity, and new process improvements were soon required. Advances in cleanroom technology and wafer cleaning, along with optimized oxidation conditions, were implemented to improve the time-zero breakdown voltage distributions. Composite oxide/nitride or SiO_2/Al_2O_3 dielectrics were also observed at that time to provide superior yield. The migration from aluminum gate electrodes to Poly-Si gate electrodes initially resulted in dramatic improvements in reliability and later allowed dual workfunction gate electrodes for optimal CMOS performance.

Constant voltage scaling in the mid- to-late '70s resulted in hot electron injection and trapping—

initially where oxide/nitride dielectrics were used and subsequently at companies using only SiO_2. In the late '70s and early '80s, the electron and hole trapping characteristics of SiO_2 were extensively studied in an attempt to optimize the processing conditions in order to produce the most stable gate dielectric. A more significant solution to the hot electron conundrum, however, came about from the lightly-doped-drain (LDD) device structure [16], rather than from improvements to the gate dielectric. Fortunately, constant voltage scaling, as described in the last several generations of the ITRS, was employed during the '90s, and the hot carrier issue has abated as power supply voltages (V_{dd}) have become much smaller, approaching I V.

As the quality of SiO_2 improved, process-induced degradation became a critical issue. Oxides grown on planar silicon substrates had very low defect densities; yet integration processes led to considerable yield loss. The "white ribbon" or "Kooi pinch" effect [17] led to a thinned gate dielectric and shorting along the LOCOS (local oxidation of silicon [18]) edges and required additional processing for its elimination. Charging damage due to implantation and dry etching were observed and found to be aggravated as the ratio of the charge collection to the discharge area (antenna ratio) increased. The radiation damage associated with e-beam evaporation of metals often precluded its use and even the sputtering of metals was seen to be potentially damaging. The control and elimination of radiation damage associated with the use of shorter wavelength lithography techniques (from the current 248 nm regime to the 193 era) is of continuing concern today.

With the scaling of SiO_2 breaching the 10 nm regime in the '90s, boron penetration through the oxide became a very serious concern. Although penetration could be reduced by avoiding hydrogen and fluorine during annealing, the incorporation of nitrogen into the oxide was seen to be more effective. Numerous approaches and processes were developed for such nitrogen incorporation, and it was soon learned that the nitrogen profile in the oxide was very important. Nitrogen at the Si-SiO_2 interface was believed to reduce strain at that interface and result in dielectrics that were more stable during hot electron injection and exhibited reduced tunnelling currents, i.e., slightly higher injection barrier heights, although too much nitrogen degrades the surface mobility. State-of-the-art dielectrics today contain considerable nitrogen, including some cases, where heavily nitrided silicon oxynitride is used to gain the additional advantage of a higher dielectric constant.

As oxides have scaled down to 2 nm, direct tunnelling (DT) of carriers through the oxide, rather than Fowler-Nordheim (FN) tunnelling, has led to excessive gate leakage currents, that increase the static power dissipation of the chip. Additionally, excessive leakage current through the gate dielectric is an ongoing reliability concern. Thus, scaling has brought us to the thickness limit for SiO_2 because of DT leakage—a fundamental, intrinsic physical phenomenon. Continued scaling of the equivalent oxide thickness (EOT), while requiring low leakage currents, will also require the use of gate dielectrics having higher dielectric constants with sufficiently large tunnelling barrier heights to suppress leakage.

THE FUTURE (2000-2020)

The 2001 edition of the ITRS has placed considerable emphasis on a high-k gate dielectric replacement for SiO_2, perhaps as soon as 2005 for low-standby power applications. The depletion layer in degenerately-doped Poly-Si gates adds about 0.3 nm to the EOT while the quantum confinement effect in silicon adds an additional 0.3 [3] nm. Thus, replacement of Poly-Si gates with dual work function metal gate electrodes will perhaps be required only a few years after 2005. An intensive global search is now in progress for new materials for these gate dielectrics and gate electrodes [19,20]. Interestingly, some of the pioneering work on field effect devices in the 1950s used barium and strontium titanate because of their high dielectric constant [21], and composite gate dielectrics using Si_3N_4 or Al_2O_3 were extensively studied in the '70s.

Encouraging results have recently been obtained on a number of candidate high-k materials, for instance HfO_2 and ZrO_2 as well as their silicates and aluminates, both as mixed oxides and as nanolaminates [19, 20]. Similarly, there is considerable interest in La_2O_3 and Y_2O_3 and their silicates and aluminates. For gate electrodes, Ta or TaN for NMOS and Ru for PMOS appear promising. Nevertheless, continued research into new materials and new processes will be necessary in order to continue scaling of MOSFET devices to the 22 nm Technology node, corresponding to L_g = 9 nm.

REFERENCES

[1] G.E. Moore, SPIE 2438, 2 (1995).
[2] H.R. Huff, G.A. Brown, L.A. Larson and R.W. Murto, ECS PV 2001-09, 263 (2001).
[3] International Technology Roadmap for Semiconductors (ITRS), (http://www.itrs.net) (2001).
[4] C.J. Frosch and L. Derrick, J. Electrochem. Soc., 104, 547 (1957). US Pat. 3,025,289 and 3,064,167 (1962).
[5] J.A. Hoerni, IRE Electron Devices Meeting, Washington, D.C., (1960).
[6] D. Kahng and M.M. Atalla, DRC, Pittsburg (1960); US Patent No. 3,102, 230 (1963).
[7] J. Kilby, US Patent No.3,138,73 (1964).
[8] R. Noyce, US Patent No. 2,981,877 (1961).
[9] B.E. Deal, M. Sklar, A.S. Grove, and E.H. Snow, J. Electrochem. Soc., 114, 266 (1967).
[10] B.E. Deal, E.L. MacKenna, and P.L. Castro, J. Electrochem. Soc., 116, 997 (1969).
[11] D.R. Kerr and D.R. Young, U.S. Patent No. 3,303,059 (1967)
[12] P.H. Robinson, and F.P. Heiman, J. Electrochem., Soc., 18, 141 (1971).
[13] Y.C. Cheng, D.R. Colton, and R.J. Kreigler, US Patent No. 3692571, (1972).
[14] C.M. Osburn, J. Electrochem. Soc., 121(6), 809 (1974).
[15] R.H. Dennard, F.H. Gaensslen, H-N Yu, V.L. Rideout, E. Bassous and A.R. LeBlanc, SC-9, 256 (1974).
[16] S. Ogura et al., IEEE Trans. Electron Dev., ED-27, 1359 (1980).
[17] E. Kooi et al., J. Electrochem. Soc., 123, 1117 (1976).
[18] J.A. Appels, E. Kooi, M.M. Pfaffem, J.J.H. Shototje, and W.H.C.G. Verkuylen, Phillips Res. Repts., 25, 118 (1970).
[19] H.R. Huff et al., Paper 1, Intl. Workshop on Gate Insulators, Tokyo, Japan, Nov. 1-2, (2001).
[20] G.D. Wilk, R.M. Wallace, and J.M. Anthony, J. Appl. Phys., 89(10), 5243 (2001).
[21] W. Shockley and J. Bardeen, Bell Syst. Tech. J., 32, 1 (1953).