Process and Device Calibration for 31/51nm NMOS/PMOS Devices fabricated by Direct Write E-Beam

H. Puchner, N. Eib, J. Kimball, M. Mirabedini, J. Haywood and S. Aronowitz

LSI Logic Corporation, 3115 Alfred Street
Santa Clara CA-95054, USA, helmut@lsil.com

ABSTRACT

In order to ensure predictability of process and device calibration tools, we manufactured CMOS devices with smallest end-of-line gate electrode dimensions of $31nm$ and $51nm$ by applying direct write e-beam lithography. A special test-chip was designed to accommodate the peculiarities of a direct write e-beam lithography process. The devices were fabricated, measured and analyzed. Process and device calibration was carried out to calibrate the threshold vs. gate length characteristics.

Keywords: Direct-write e-beam system, deep-submicron devices, process simulation, device calibration.

1 Introduction

It is important for the TCAD tools to maintain their predictability and accuracy as device dimensions shrink. State-of-the-art devices show increasing small device geometry effects. The most important ones from the viewpoint of TCAD are the reverse short channel effect, reverse narrow width effect, drain induced barrier lowering, gate induced drain leakage, and junction leakage currents. All these effects play an important role when simulating the electrical characteristics of deep sub-micron devices. Devices are also starting to exhibit significant quantum mechanical effects, such as tunneling effects and inversion layer quantization. The gate tunneling current causes a significant contribution to the overall leakage current for gate oxides below 20Å. Therefore, it is extremely important to accurately simulate leakage currents. Modern submicron devices also require an additional large angle pocket implant to prevent punchthrough at short channel lengths. The pocket implant therefore adds to the total channel dose and can be considered a major dopant source for the "reverse short channel effect" (RSCE)[1], [2]. As the pocket implant dose increases, we observe a drastically increased RSCE. TCAD tools must be able to capture this dramatic rise in threshold voltage by adjusting the interstitial recombination rate at the Si/SiO_2-interface to achieve the right amount of channel dopant redistribution on one hand, but not to overdiffuse the source/drain junctions in the lateral as well as vertical direction on the other hand. Especially, the lateral diffusion is critical in modeling the roll-off behavior of the threshold voltage vs. gate length characteristic. In order to estimate the device behavior at deep submicron device dimensions ($< 100nm$) one has to be able to calibrate the simulation tools to real silicon experimental data beforehand. Only in this case it is possible to sustain predictability of the process and device simulation tools. Since the lithography equipment for printing e.g. $50nm$ transistor structures in a production environment has not been fully developed yet, we employed direct write e-beam lithography to define deep sub-micron transistor gates.

2 Direct Write E-beam

Several different technologies are considered as possible solutions for the next generation lithography tools. X-ray lithography exhibits the drawback of feature size masks [3], which limits the capability of mask defect detection and mask repair. E-beam direct write can only achieve reasonable write times if multiple beams are be employed. The most promising method seems to be SCALPEL [4], which combines the high resolution inherent in electron beam lithography with the throughput of projection systems. A mask membrane scattering mechanism is used to achieve high resolution images. However, for low throughput applications, such as prototyping, it is still applicable to employ direct write E-beam systems. We used a Leica $100keV$ direct write E-beam system at the Nanofabrication Facility at Cornell University to print polysilicon gate structures down to $48nm$. Extensive electron beam, dose, and pixel size calibrations were carried out to ensure reliable printing of the minimum features. The write time could be further reduced by only sampling every other bit of the mask subfield resulting in a minimum pixel size of $16nm$. Figure 1 shows the dependence of the final polysilicon dimension on the e-beam write dosage. The dose was chosen to achieve minimum feature sizes of $48nm$. Optimization of the test chip layout was carried out to reduce the poly area and, hence, the die write time. The final write time could be improved to about $600\mu m^2/s$, totaling about 4min. per die for the selected test chip. A total number of 16 dice per wafer were exposed and processed at the Nanofabrication Facility.

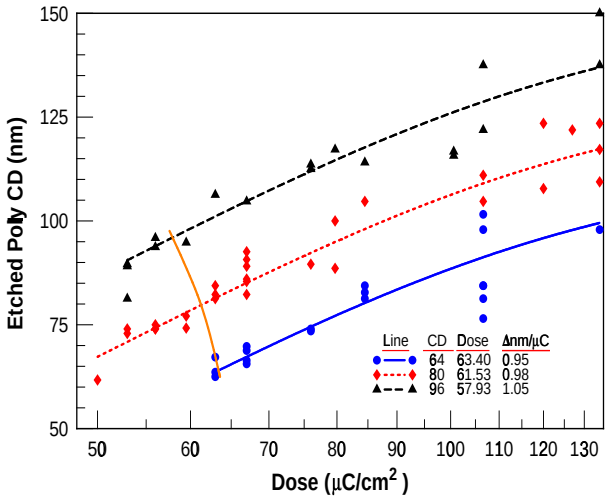

Figure 1: etched polysilicon CD dimension dependence on e-beam write dosage. Dose is optimized to minimum feature sizes.

All other mask layers and processing was carried out at LSI Logic's manufacturing facilities.

3 Transistor Design

Deep submicron transistors require a careful design of source/drain junctions as well as channel dopings. The threshold voltage roll-off has to be controlled for short channel devices by introducing pocket implants. Other important design parameters are the LDD implant dose and energy as well as spacer width and poly gate electrode thickness. The poly gate electrode thickness has to be aligned to the source/drain junctions to avoid counterdoping of the channel by the source/drain implants. The most critical building block of a very deep submicron technology is the gate dielectric layer. The continuous down-scaling trend in device dimensions drives the further decrease of the gate oxide thickness. Deep sub-micron CMOS technologies require gate oxide thicknesses well below 20Å [5]. One of the most severe problems caused by the employment of thin gate oxide layers is the boron penetration from the heavily doped p+ polysilicon gate trough the underlying gate oxide and into the channel region of the PMOS transistor. This leads to the decrease in threshold voltage down to a surface source-drain punchthrough [6]. Boron penetration also degrades the device reliability by generation of defects in the gate oxide. One possible solution to suppress boron penetration is gate oxide hardening by incorporation of nitrogen into the gate oxide [7], [8]. Different approaches to nitrogen incorporation into the gate oxide layer have been developed. Among them there are formation of a nitrogen atom monolayer at the Si/SiO_2 interface between gate oxide and substrate[8], incorporation of nitrogen atoms in gate oxide bulk, and formation of a thin heavily nitrogen doped layer at the top of the gate oxide [9]. Each of these approaches ad-

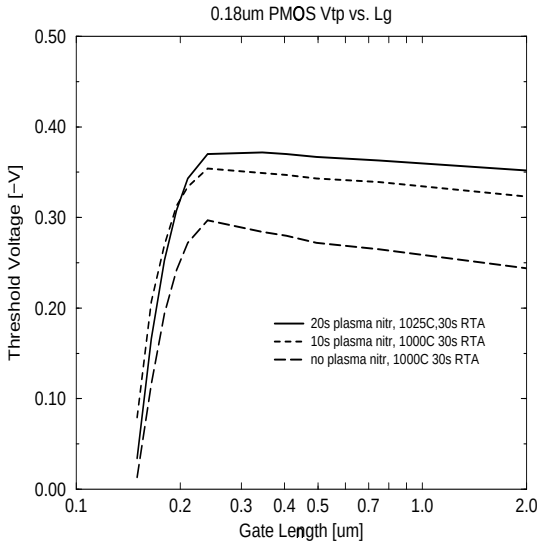

Figure 2: Threshold vs. gate length curves for different plasma nitridation times. Best results are obtained for 20s exposure time of the gate oxide to the nitrogen plasma. Even when the sample was annealed at higher temperature no sign of boron penetration was found.

dresses a specific problem with respect to device reliability and performance. For example, a monolayer nitrogen incorporation at the Si/SiO_2 interface reduces the interface roughness and therefore improves the high field mobility and reduces defect generation under hot carrier stressing. Bulk nitridation reduces the electrical oxide thickness by maintaining the same optical thickness resulting in the employment of thicker gate oxides and the advantages associated with it, e.g. lower direct tunneling current. Top surface nitridation is the most effective method of avoiding boron penetration. It offers several advantages compared to the creation of a nitrogen diffusion barrier at the bottom of the gate oxide. Depending on the method used to create the barrier the nitrogen at the Si/SiO_2 interface will increase the interface trap density causing possible mobility degradation. Additionally, the boron trapped in the bulk of the gate oxide increases the electron trap density [10] throughout the gate oxide. The latter will increase the defect generation under Fowler-Nordheim and hot carrier stressing[11]. Top surface nitridation of the gate oxide is almost free from these problems. Conventional high temperature oxidation in N_2O or NO ambient results in a nitrogen incorporation at the Si/SiO_2 interface of less than 5% [12]. Two different techniques allow much higher level of nitrogen incorporation at the gate oxide top surface: deposition of an ultra-thin silicon nitride layer on the top of the gate oxide [11] and the ultra-low energy ion bombardment from a nitrogen plasma source [9]. The strength of the diffusion barrier will depend on the nitrogen concentration as well as on

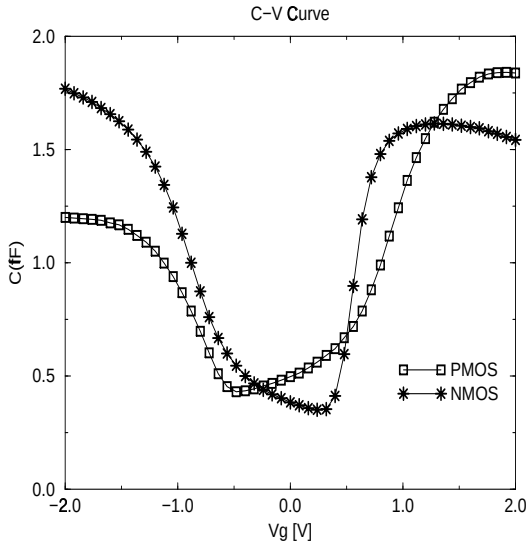

Figure 3: C-V curves for NMOS and PMOS devices. NMOS data were extracted from samples which received plasma nitridation, where PMOS samples received an additional pre-gate nitrogen implant and etchback before nitridation.

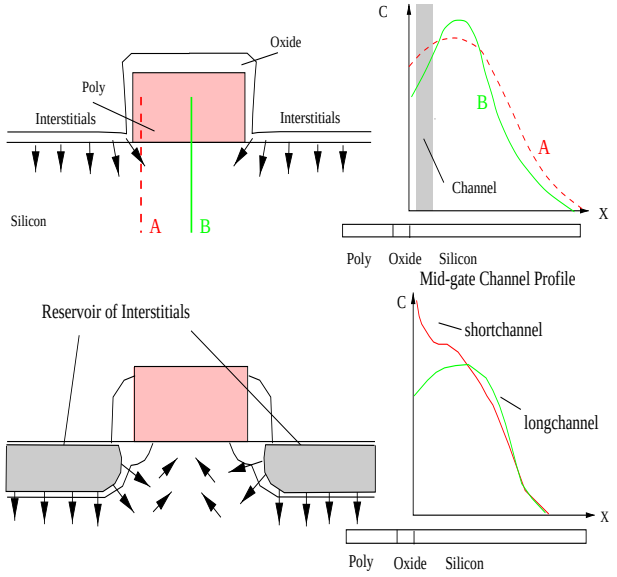

Figure 4: Redistribution of the channel dopants caused by the oxidation injected interstitials (top) and source/drain point-defects during the source/drain annealing (bottom).

the distribution of the incorporated nitrogen within gate oxide.

We chose ion bombardment from a nitrogen plasma to achieve optimum nitridation conditions. Several short-loop experiments were conducted to established sufficient surface nitridation. The wafer chuck was grounded to avoid etching of the surface by the nitrogen ions. Figure 2 shows the threshold voltage vs. gate length characteristic for different transistors exposed to different plasma nitridation conditions. Where the control sample shows severe signs of penetration, the plasma nitrided gate oxides can withstand penetration even at 1025° C annealing temperature. Intentionally, the target gate oxide thickness for this experiment was chosen to be $15-17$Å, however, due to mis-processing the final physical oxide thickness was measured by TEM cross-sections to be 31Å for the NMOS device and 25Å for the PMOS device, which has undergone different gate oxide treatment. The inversion oxide thickness was extracted from C-V curves (see Fig. 3) to be 30Å for NMOS devices and 20.6Å for PMOS devices.

4 Process- and Device Calibration

Several explanations have been proposed for the unexpected increase in the threshold voltage of MOSFETs when the channel length is decreased. Jacobs et al. [1] assumed position dependent interface charges injected by interstitials into the Si/SiO_2-interface. Generally, the RSCE can only be explained by assuming different channel doping conditions for long and short channel devices. Consequently, for a conventional CMOS process flow all process steps prior to the formation of the gate electrode are the same for long and short channel transistors and therefore not important. The first process step impacting the RSCE is the polysilicon gate electrode reoxidation. The polysilicon reoxidation serves several purposes. It reoxidizes the thin gate oxide on top of the source/drain regions. It smoothes out the polysilicon sidewalls, which are usually rough from the previous poly etch process step. The polysilicon reoxidation also improves reliability of the device by forming small bird's beaks at the gate corners. This thicker corner gate oxide reduces the electric field at the corner regions as well as the gate overlap capacitance. Unfortunately, the oxidation process injects interstitials into the channel region. These injected interstitials result in an enhanced diffusion of dopants near the interface, the so-called *oxidation enhanced diffusion* (OED)[13]. However the mid-channel dopants are nearly unaffected by point-defects injected at the gate corners because the point-defects have already recombined at the Si/SiO_2-interface or in the bulk. The OED causes different surface concentrations in the channel for long and short channel devices. The major source for RSCE is the redistribution during the source/drain annealing. The more dopants that are present in the channel region the more likely the dopants will be redistributed by the point-defects introduced by the source/drain ion implantation step. The interstitials generated by the source/drain implantation are drifting towards the gate oxide surface as well as the bulk during the annealing process. By passing through the channel region they interact with the chan-

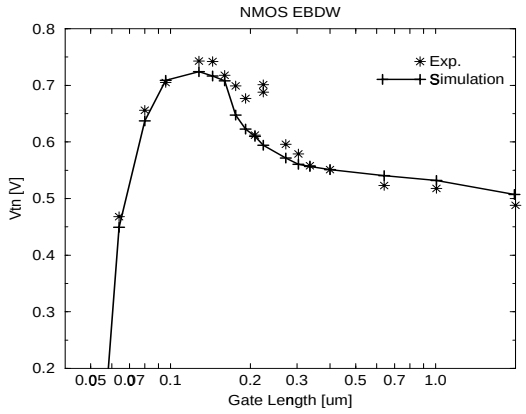

Figure 5: Comparison of experimental and simulation results for dual pocket threshold voltage vs. gate length dependence of NMOS devices with second pocket implant dose of $7.5 \cdot 10^{12} cm^{-2}$.

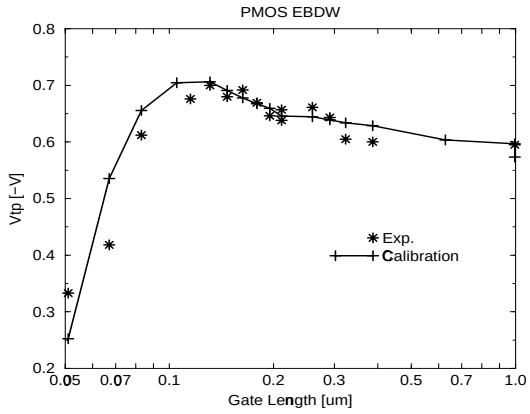

Figure 6: Comparison of experimental and simulation results for dual pocket threshold voltage vs. gate length dependence of NMOS devices with second pocket implant dose of $1 \cdot 10^{13} cm^{-2}$.

nel dopants and push them towards the surface. The flux of interstitials towards the surface depends now on the capability of the Si/SiO_2-interface to capture interstitials. Figure 4 illustrates the interaction of the point-defects with the channel dopings. The interstitial interface recombination rate in combination with the generated point-defects (+1-model) and the source/drain annealing conditions are the most effective parameters in calibrating the simulator input deck towards the experimental data. Additionally, the interstitial evaporation rate at the surface must be taken into account, because it reduces the amount of available mobile interstitials. Since two-dimensional dopant profiles are commonly not available, electrical characteristics have to be calibrated instead. Usually, the threshold voltage vs. gate length characteristic needs to be calibrated to ensure appropriate capture of short and long channel effects.

5 Simulation Results

TSUPREM4 was used for the thermal and topological process simulation steps [14], where the device simulations were carried out by Medici [14] to obtain the threshold voltage vs. gate length characteristics. By applying the above described calibration parameter set simulation results were obtained for NMOS and PMOS devices of different gate length. Figure 5 and Figure 6 show the threshold voltage vs. gate length characteristics for NMOS and PMOS devices, respectively. Good agreement is found when comparison is made with experimental data.

6 Conclusions

We successfully manufactured transistor test structures down to $31nm$ end-of-line ($48nm$ printed) gate electrodes using direct write e-beam technology. It was possible to calibrate NMOS and PMOS devices with current process and device simulation models. There was no major limitation or roadblock encountered for extending the simulation models down to the 30nm transistor technology range.

REFERENCES

[1] H. Jacobs et al., IEDM Techn. Digest, pp. 307–310, 1993.
[2] C. Rafferty et al., IEDM Techn. Digest, pp.311–314, 1993.
[3] N. Mizusawa et al., 2000 Int. Conf. on Microprocesses and Nanotechnology, pp.108-109, 2000.
[4] K. Tokunaga et al., VLSI Techn., 2000 Digest of Techn. Papers, pp. 54-55, 2000.
[5] G. Lucovsky, IBM J.Res.Develop., 43 (3), 301 (1999)
[6] J. Pfiester et al., IEEE Elec. Dev. Lett. 11, 247 (1990).
[7] D. Lee et al., J. Vac. Sci. Technol. A 13, 607 (1995).
[8] G. Lucovsky et al., J. Vac. Sci. Technol. A 14, 2832 (1996).
[9] S. Hattangady et al., Appl. Phys. Lett. 66, 3495 (1995).
[10] J. Pfiester et al., IEEE Trans. Elec. Dev. 37, 1842 (1990).
[11] Y. Wu et al., IEEE 98CH36173, Proceedings of 36th Annual Intern. Reliability Physics Symp. 70 (1998).
[12] Z. Ma et al., IEEE Elec. Dev. Lett., 15, 109 (1994).
[13] K. Taniguchi et al., J. Electr. Chem. Soc., 127(10), pp. 2243–2248, 1980
[14] Avant! Corp, TSUPREM4 and Medici Users Guide, Fremont CA.