Gate technology for 70 nm metal–oxide–semiconductor field-effect transistors with ultrathin (<2 nm) oxides

D. Tennant[a)]
Bell Laboratories, Holmdel, New Jersey 07733

F. Klemens, T. Sorsch, F. Baumann, G. Timp, N. Layadi, A. Kornblit, B. J. Sapjeta, J. Rosamilia, T. Boone, B. Weir, and P. Silverman
Bell Laboratories, Murray Hill, New Jersey 07974

(Received 12 June 1997; accepted 8 August 1997)

Results are described for a gate level technology module developed to produce metal–oxide–semiconductor transistors with physical gate lengths of 70 nm and below. Lithography is performed by direct write e-beam lithography (EBL) using a thermal field-emission EBL system in SAL 601 resist. Critical dimension (CD) control, as measured by several methods, is found to depend not only on dose control but also on writing parameters such as pixel spacing. The pattern transfer using a silicon dioxide hard mask is shown to exhibit a trade-off between anisotropy and selectivity. Transmission electron microscopy cross sections reveal that two atomic layers are removed even when the gate oxide stopping layer is completely intact. We report results for gate lengths down to 60 nm with edge roughness on the order of 5 nm, within the acceptable limits for threshold requirements, while stopping the etch process on oxides as thin as 1.2 nm. © *1997 American Vacuum Society*. [S0734-211X(97)07306-X]

I. INTRODUCTION

For the last 30 years, one of the principal means for improving integrated circuit performance has been miniaturization of the transistors and wires that comprise it. For this trend to continue, several vexing problems have yet to be resolved, however. For example, metal–oxide–semiconductor (MOS) transistor technology currently under development uses 0.25 μm design rules with a 6 nm thick gate oxide and a 3.3 V power supply. Employing a constant field scaling scenario to predict the characteristics of smaller metal–oxide–semiconductor field-effect transistors (MOSFETs) well into the future, we anticipate that a 70 nm channel length transistor will require an ultrathin gate oxide less than 1.7 nm thick and a power supply of less than 1 V. Simulations using PROPHET[1] and PADRE[2] indicate that dimensional control for critical level lithography will be severe, and interfaces smooth to nearly atomic precision will be required to produce these transistors. For example, at a supply voltage of 1 V, 7 nm variations in the channel lithography would give rise to a 0.1 V change in the threshold voltage and, hence, a 50% increase in the delay of an inverter, while 0.2 nm variations in the oxide thickness would increase the gate tunneling current by a factor of 10. Therefore, the tighter the control on features, the more aggressive a given high-performance circuit design can be.

A key element in the design of a 70 nm gate length MOSFET is, therefore, definition of the gate stack. All aspects of the preparation of the active channel region of the transistor have important consequences for the expected performance. For example, the cleaning method used on the improved epi-wafers can drastically effect surface roughness and, thus, the quality of the Si/SiO_2 interface.[3] In this article, we will emphasize the lithographic and pattern transfer results relating to this scaled silicon initiative but will also raise other issues and findings, which are related to the expected performance of the transistors being developed.

There are two candidate *n*-channel metal–oxide semiconductor (NMOS) device testers that are used to implement this gate technology module. The first is a single level tester MOSFET, which is designed with one lithographic level and, therefore, can more rapidly provide dc characteristics to both verify device simulation models and to provide engineering feedback for critical process steps. The second is an end-to-end NMOS test lot, which will provide ac electrical characteristics and circuit performance data.

II. FABRICATION

We have been exploring the utility of a gate stack consisting of 100 nm of a silicon dioxide hard mask over 80 nm of WSi_x on 100 nm of polycrystalline silicon on gate oxides ranging in thickness from 4 to 1.2 nm. The epitaxial silicon wafers are prepared in-house and have a rms surface roughness measured by atomic force microscopy (AFM) of less than 0.08 nm. For the tester MOSFETs, the wafers are received unpatterned prior to gate level lithography. The end-to-end wafers have had several lithography and process steps prior to application of the gate stack, among these, a deep dry etch step in which e-beam alignment marks are formed. In either wafer series, the gate stack process is the same, although the full NMOS process restricts the thin gate oxide growth to selected areas in order to achieve isolation.

There are at least three processing steps that can adversely affect the gate oxide interface quality: (1) the sacrificial oxide growth; (2) the cleaning prior to oxidation; and (3) the

a)Electronic mail: dmt@bell-labs.com

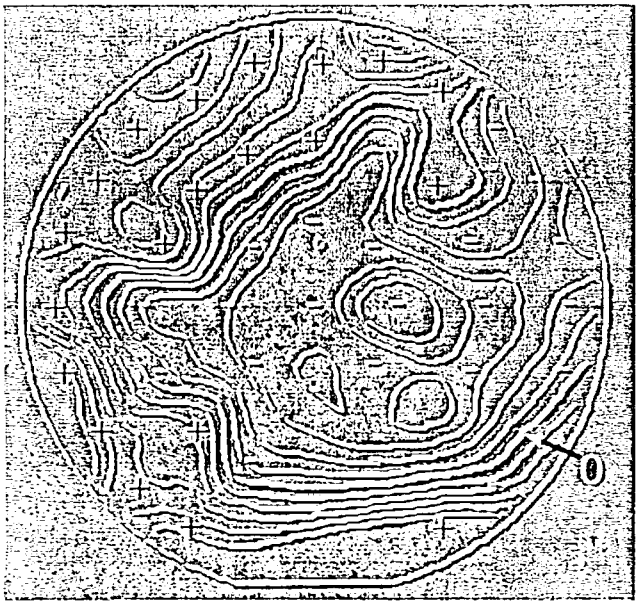

FIG. 1. Uniformity of the ultrathin gate oxide grown on a 150 mm wafer. The mean thickness is 1.545 nm with a standard deviation of 0.024 nm. The contour line are in steps of 0.01 nm. The minimum thickness is 1.499 nm and the maximum is 1.592 nm. The measurement are made using a scanning ellipsometer.

gate oxide growth. The key issues in each case are to protect the silicon substrate from roughening while removing metals, particles, and other surface contaminates, and to produce a uniform oxide of a specified thickness.

A. Gate stack formation

We have found that using an *in situ* vapor phase UV/Cl_2 cleaning strategy in conjunction with rapid thermal oxidation, the interface roughness of the starting substrates can be maintained throughout the process flow. This process was found to be substantially improved over conventional wet chemical cleans with similar bulk substrates.[3] Vapor-phase processing was carried out at 100 °C in 10 Torr of chlorine gas, activated by ultraviolet radiation for an interval ranging from 7 to 60 s. Prior to the *in situ* treatment, the native oxide was stripped by immersion into a 15:1 H_2O:HF solution for 5 s.

After cleaning, the wafers enter the rapid thermal oxidation reactor of an IntegraOne cluster tool. The wafers subjected to UV/Cl_2 treatment enter the reactor without exposure to the ambient in a vacuum of 10^{-7} Torr, while the elapsed time between the termination of a wet clean and introduction into the reactor was held to less than 5 min. For oxide growth, the wafers are initially ramped to 1000 °C in a partially oxidizing atmosphere over a 30 s interval. After stabilizing the temperature at 1000 °C, oxidation occurs in pure oxygen in a pressure ranging from 1 to 500 Torr. Following oxidation, the wafer is annealed at 930 °C for 30 s in 99.999% argon. The entire oxidation/anneal cycle may be as long as 2 min in duration, depending on the desired oxide thickness, and the pressure during oxidation. Figure 1 is an example of the uniformity observed using a scanning ThermaWave Optiprobe ellipsometer to characterize the thickness variation across a 150 mm wafer. In this example, the maximum departure from the mean of 1.545 is less than 0.05 nm over the entire wafer and within the 0.03 nm requirement for our work if we limit devices to the central 100 mm region.

AFM was used to characterize the silicon surface roughness, in tapping mode with sharpened silicon tips, following a previously reported procedure.[4] Roughness measurements were obtained for Si substrates directly after cleaning with no sample preparation. For oxidized wafers, the roughness of the oxide surface was obtained directly, and then the interface roughness was measured after etching the oxide layer in a 1:1 solution of $HF:H_2O$.

The UV/Cl_2 cleaning has been associated with roughening of the Si surface. Ma and Green[5] indicated that roughening might be minimized with very short UV/Cl_2 exposure time, without compromising reliability. In our experience, the interface did, in fact, retain much of the character of the substrate and was measured to be smoother (0.08 nm rms) than those obtained with standard wet cleaning (0.13–0.16 nm rms).

The UV/Cl_2 processing consistently results in a capacitor with lower leakage than wet chemical cleans as discussed by Sapjeta *et al.*[3] However, we cannot unambiguously discriminate between the various cleans and the corresponding roughness based on these characteristics alone, because of subnanometer scale nonuniformities in the oxide thickness found across the wafer, and because of our inability to accurately and precisely assess the thickness with ellipsometry and transmission electron microscopy (TEM) to better than 0.13 nm. However, a factor of 5 improvement in Q_{BD}, a measure of charge to breakdown (used to evaluate reliability), was observed with the UV/Cl_2 cleaning strategy, which we attribute to a smoother Si/SiO_2 interface. No metal contamination was detected with total x ray reflection fluorescence (TXRF) and secondary-ion-mass spectroscopy (SIMS) analysis on any of these wafers (e.g., $Fe<0.8\times10^{10}\,cm^{-2}$). We generally observe silicon band tunneling at low voltages, and the leakage current is generally lower, for the same ellipsometric and TEM thickness, than that reported by Schuegraff *et al.*[6] and Momose *et al.*[7]

The remaining gate stack is next deposited using standard methods. The 100 nm layer of polysilicon is deposited by chemical vapor deposition (CVD) at 550 °C using silane and is *in situ* doped with phosphorus using PH_3. The as-deposited film is amorphous but is annealed to form polycrystalline material. The 80 nm WSi_x ($x\sim2.7$) is dc magnetron sputter deposited from a single target. The final layer of the gate stack is the 100 nm hard mask, which is a conformal, low-temperature low-pressure CVD oxide deposited using a TEOS (tetraethoxysilane) decomposition process.

B. E-beam lithography

The gates with lengths in the range 70–250 nm are patterned via e-beam lithography with a JEOL JBX 6000FS

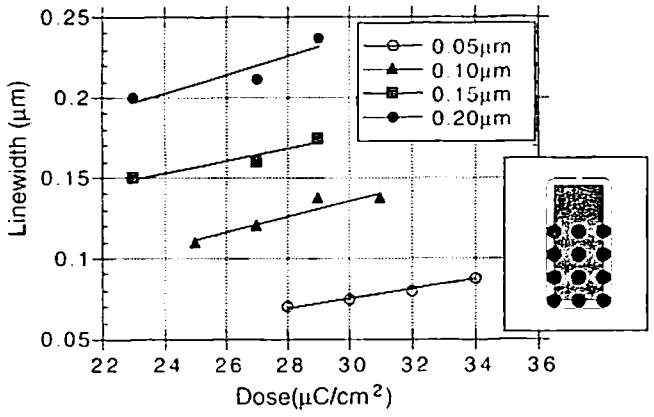

FIG. 2. Linewidth dependence on area dose for 50 kV electron-beam lithography for different coded linewidths. This data is used to optimize the gate doses. The inset illustrates why we expect the developed profile to be wider than the coded linewidth. Under the exposure conditions typically used, we observe about a 20 nm increase over the coded size.

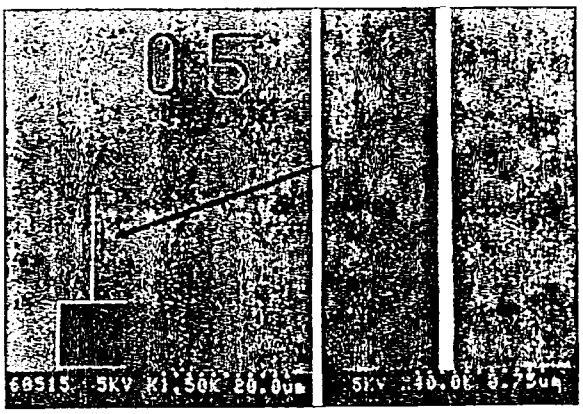

FIG. 3. SEM of the gate region of a 0.05 μm coded discrete n-channel MOSFET in 0.28 μm thick SAL601 ER4 resist patterned on the gate stack and prior to pattern transfer. The enlarged photograph demonstrates the smooth edges needed to control threshold variations.

thermal field-emission (TFE) system. This step is not only challenging because of the 70 nm linewidths in the minimum gate length devices but because of the concurrent requirements to: (1) produce high-resolution lithography with gate length variations along the gate width of less than 5 nm; (2) provide a process that minimizes artifacts related to circuit topography; (3) provide sufficient throughput during direct write to allow for both systematic evaluation of gate length dependent performance and permit splitting of wafer lots to bracket critical process parameters (e.g., low-energy implantation parameters); and (4) produce gate-to-gate, chip-to-chip and wafer-to-wafer CD control in order to properly interpret causal performance effects.

To develop a process to meet the outlined requirements we selected Shipley SAL 601, a negative, chemically amplified resist for its low sensitivity but high resolution. Figure 2 plots the linewidth variation as a function of dose for various size coded lines for a standard prebake condition of 85 °C and postexposure bake (PEB) of 105 °C. All bakes were performed on vacuum hot plates. The data in Fig. 2 is used to assign optimum doses to various size critical features (e.g., gates) as a simple dose-modulation proximity effect correction. As illustrated in the inset, the expected linewidth is typically larger (by about 20 nm) than the coded linewidth since the writing method places the beam on the perimeter of the coded figure. The expected broadening is one beamwidth (~15 nm) plus the *fast*-secondary electron range.[8] Typical linewidth variations are observed by scanning electron microscopy (SEM) such as those shown in Fig. 3. More quantitative measurements were obtained from a KLA 8100 CD measurement system. For linewidths greater than 70 nm, typical maximum variations were found to be 5 nm.

Other factors were explored such as beam current effects and minimum address size in the coded linewidths. Surprisingly, the use of a 25 nm address to fill the area was found to significantly reduce the linewidth of a 150 nm feature at a given dose over that of the same feature exposed with a 50 nm address, even at the same beam diameter, current, focus, PEB, and development conditions. Moreover, the linewidth was rather insensitive to the beam current (and, therefore, the beam diameter) over the range from 350 pA to 1.4 nA. One would like to take advantage of these observations and reap the throughput benefit with no penalty in linewidth by using the higher current at the smaller address. However, the deflector speed in our e-beam system limited the current to about 500 pA for the 25 nm pixel size. Clearly, higher deflection speeds are needed to fully utilize the high current density available from TFE sources.

Figure 4 is an atomic force micrograph of the gate lithography obtained from the resist profile written in 350 nm of SAL-601 ER4 resist using an exposure base dose (i.e., the dose given to the large pad region) of 15 $\mu C/cm^2$, with proximity corrected features receiving an elevated dose. This image serves to illustrate the demands that topography adds to

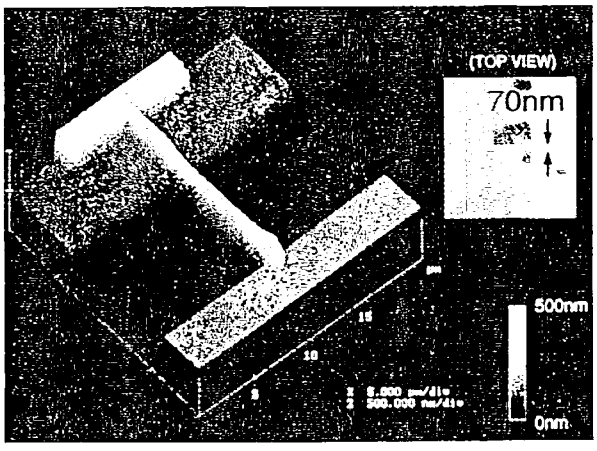

FIG. 4. Atomic force micrograph of a 0.07 μm gate pattern in 0.35 μm of SAL 601 ER4 resist. The color image illustrates the demanding topography required in the transition regions between the polybuffered LOCOS (PBL) region and the active thin oxide region. The inset shows a top-down view of the same image which revealed a systematic narrowing of the gates in the region of the thick oxide. This was cured by increasing the dose in that area.

TABLE I. Etching parameters for hard-mask RIE in Applied Materials AMI 5000 magnetically enhanced etcher.

Etch step	Pressure (mT)	Power (W)	Field (G)	CHF_3 (sccm)	SF_6 (sccm)	CF_4 (sccm)	Ar (sccm)
Main	15	625	60	65	3	...	...
Over-etch	85	625	60	30	...	4	60

the task of defining a 70 nm feature in the gate level lithography. A careful study of these and similar data revealed a slight narrowing in the gate in the region near the pad, which is written on the local oxidation of silicon (LOCOS) region. In subsequent exposures, a dose adjustment to this area was made to correct this. After various trials, a nominal resist thickness of 200 nm of SAL 601-ER2 was determined to provide a good compromise between resolution and etch resistance, thus, limiting the maximum aspect ratio in the resist to 3.

C. Pattern transfer

The resist pattern is transferred using reactive ion etching (RIE) which, together with the lithography, determines the physical device length which is crucial to the electrical characteristics of the transistor. Therefore, the cross-sectional profile of the gate stack should be close to vertical and the etch must have sufficient selectivity to stop on the ultrathin gate oxide. To achieve this, the TEOS hard mask is etched in CHF_3/SF_6, which provides excellent fidelity to the resist. The etch into the hard mask is performed in an Applied Materials AMI 5000 magnetically enhanced RIE system using the parameters found in Table I. The main etch end point detection is determined using an optical emission type signal. Following the oxide etch, the resist is removed.

The remaining gate stack RIE is performed in a LAM Research TCP 9400SE transformer coupled etcher. The tungsten silicide is etched using Cl_2 in a 20% He/O_2 background under the conditions found in Table II. The gate structure is completed by transferring the pattern into polycrystalline silicon using HBr in Cl_2 followed by an overetch in which the Cl_2 is replaced with He and a He/O_2 mixture in order to stop on the ultrathin gate oxide with very high selectivity. Optical emission end-point detection is again used to determine a consistent stopping point. The dc bias voltage is not monitored in these commercial high-density plasma systems, however, at the tabulated powers we speculate that the ion energies are below 100 eV and may be substantially below this.

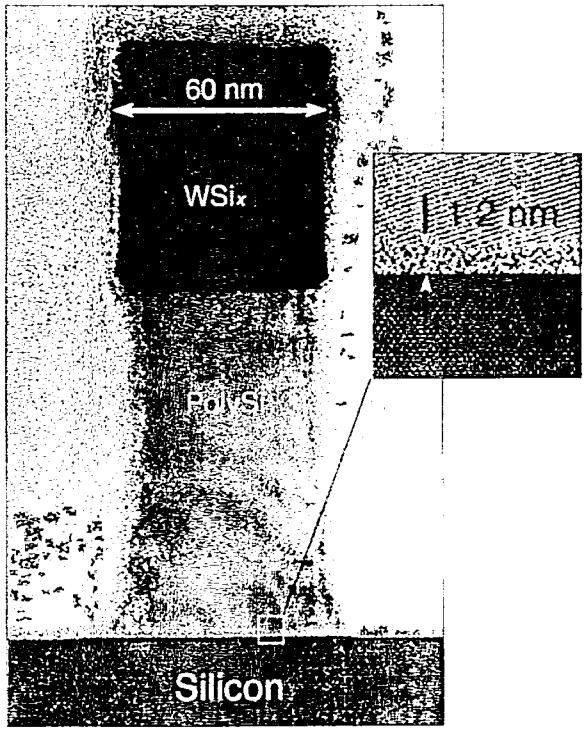

FIG. 5. Transmission electron micrograph of a cross-sectioned 60 nm long gate stack structure with a 1.2 nm gate oxide patterned by electron-beam lithography and high-density plasma etching.

III. ANALYSIS

Figure 5 shows a TEM of the 60 nm gate profile resulting from this etch. We note that the gate is nearly perfectly transferred, but undersized by about 5 nm when compared with the original resist pattern (as measured by the SEM-based CD instrument). Using the silicon lattice as a self-calibration, the gate oxide thickness can be determined. The oxide in this device is 1.2 nm±0.13 nm, as seen in the inset of Fig. 5.

While atomic force microscopy and transmission electron microscopy confirm that the gate oxide resists the etch, we consistently noted two phenomena, which can be observed in Fig. 6. First, the oxide appears to thicken in the regions unprotected by the gate stack. The increase appears to be in the range 1.2–3.3 nm added to the top of the gate oxide, sug-

TABLE II. Etching parameters for gate stack RIE in a LAM Research TCP 9400SE transformer coupled etcher.

Etch step	Pressure (mT)	TCP power (W)	Chuck power (W)	Cl_2 (sccm)	HBr (sccm)	He (sccm)	He/O_2 (5:1) (sccm)
Break-through	5	250	200	80	...	...	...
Tungsten silicide	2	250	200	50	...	...	10
Polysilicon	2	250	180	30	150	...	...
Overetch	60	300	200	...	100	200	10

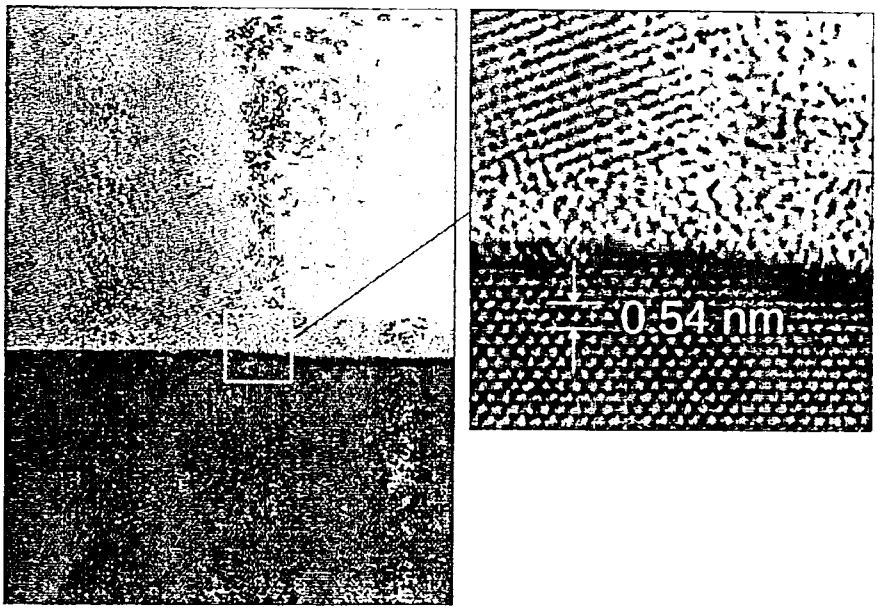

FIG. 6. TEM of the "foot" region of the etched gate structure in Fig. 5. A thickening of the oxide layer is observed in the region exposed to the etch plasma. About two atomic layers of silicon appear consumed during the etch process.

gesting that it may be due in part to a redeposition process. More study is needed to confirm this, however. Second, there is some evidence beyond the edge of the gate mask of minute consumption of the underlying silicon in the overetch. The high-resolution TEM in Fig. 6 allows us to estimate the consumption in this region to be about two silicon monolayers. Etch results on slightly thicker oxides (2.5 nm) were nearly perfect showing no such penetration. In addition, when the main poly-Si etch was stopped sooner and the etch completed using the overetch recipe, a similar intact silicon layer was observed. Under these "less aggressive" conditions, however, a small foot was observed to develop at the base of the gate stack slightly lengthening the physical gate.

IV. EVALUATION DEVICES: DISCUSSION AND SIMULATIONS

In order to more rapidly evaluate the validity of our process and electrical performance simulation parameters and models, the simplified single level FET structure shown in Fig. 7 was designed. The figure shows a SEM image of the SAL 601 resist pattern on silicon, which is representative of 40 device quality 150 mm wafers direct written on the gate stack materials and, subsequently, processed. The outside dimensions of the gate pad region are 320 μm×320 μm. While the large gate pad area prevents meaningful high-

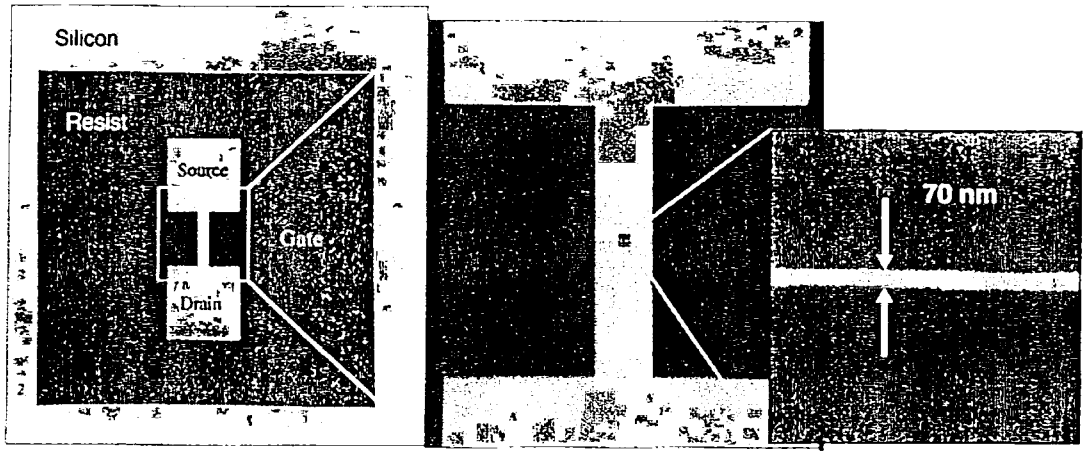

FIG. 7. SEM of a reduced mask level transistor being fabricated to more rapidly evaluate the design of the gate formation process. The large difference in W/L between the short channel FET and the large-area transistor causes the short channel device to dominate the conductance.

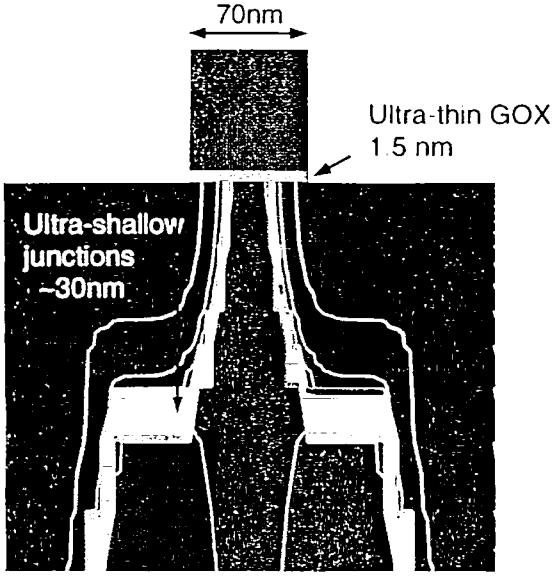

FIG. 8. PROPHET process simulation of the device shown in Fig. 7. The red color denotes *p* type and the blue denotes *n* type. In the *n*-type region the three yellow contours delineate the regions with different decades of doping levels starting at 10^{19} cm^{-3}. In the *p*-type region, the single contour easily visible in the plot you have delineates a contour of 10^{18} cm^{-3}. The doping near the channel is approximately 10^{17} cm^{-3}. The simulation estimates the effective channel length to be about 27.5 nm and the LDD junction depths to be about 30 nm. This device is expected to exhibit a g_m of 1.85 S/mm.

frequency measurements, the structure can be very useful in providing estimates of dc electrical performance, contact characteristics, etc.

The pattern was e-beam written using a simple sleeving method of proximity correction in which various sections of a several micrometer wide outline of the pattern are assigned doses independently from the interior large areas. This allowed for well-defined edges and corners over the entire structure as well as the short gate. The design of this structure allows for multiple conduction paths, but has been modeled to ensure that owing to the large differences in the $W_g:L_g$ (width-to-length) ratio, the conduction through the short gate area swamps any parallel conduction path.

While a full description of the device process beyond the gate module is deferred to future reports, PROPHET simulations were performed for this device using nominal target implant and anneal schedules. The resulting structure is shown in Fig. 8. The simulation includes models for transient enhanced diffusion, ultra-low-energy implants (<4 kV), the ultrathin (<2 nm) gate oxides, recessed LOCOS and poly-buffered LOCOS (PBL) device isolation, and an 80 nm dielectric sidewall deposition for the deep implant. The results reveal an effective channel length of 30 nm and a junction depth in the low doped drain (LDD) region of 30 nm. The LDD region is found to extend beneath the physical gate by about 15 nm. This suggests that the two monolayer step observed after the RIE pattern transfer of the gate should not have an observable effect on the electrical performance of these devices.

The electrical performance of this structure was then simulated using PADRE and included models for drift diffusion, energy balance, field-dependent mobility, tunneling, and Lentz/Klassen mobility models to describe the effects of surface phonon scattering and surface roughness. The predicted transconductance, g_m, is 1.85 S/mm, the subthreshold slope, $S=92$ mV/dec. The estimated value of $I_{on}=1.7$ mA/μm at a drain voltage of 1.0 V, and $I_{off}=0.95$ μA/μm at a gate voltage=0 V. While these performance goals are an exciting prospect, they are only simulations and we look forward to having completed devices in the near future.

V. CONCLUSIONS

We report a 70 nm gate technology module, which meets the initial R&D requirements of CD, gate variation, ultrathin oxide growth, and pattern transfer fidelity and selectivity. We have applied this technology in a two tiered approach to making NMOS FETs. To more rapidly gain experience with the modeling validity in this regime, we have designed a structure, which exploits width/length ratio differences between a short gate region and other conduction paths. Best efforts to date for gate lithography have realized 65 nm gate lengths with edge roughness in the e-beam written gates on the order of 5 nm, within acceptable limits for threshold requirements.

Pattern transfer was nearly perfectly vertical but an overall narrowing by an estimated 5 nm was observed, resulting in a 60 nm gate formation. The most critical component of the etch is stopping on gate oxides as thin as 1.2 nm. TEM analysis of the etched gate stacks reveal that for all cases the unprotected areas beside the gates grow in thickness in excess of 1.2 nm. When an aggressive etch recipe is used to optimize the fidelity of the patterning, we find for oxides thinner than 2 nm that about 2 atomic layers of the silicon are consumed below the surface of the oxide in the region not protected by the gate mask. For thicker oxides (e.g., 2.5 nm) this is not observed. When a less aggressive etch recipe is used on the ultrathin oxide wafers, we observe stoppage with no measurable consumption of the silicon below the oxide. This improvement comes at the price of degraded anisotropy, however.

We have implemented this gate process in devices, which have been simulated to exhibit transconductances in the range of 1–2 S/mm. We hope to have confirming electrical measurements in the near future.

ACKNOWLEDGMENTS

The authors wish thank R. Kasica and K. Feder who helped with the e-beam lithography. The authors also wish to thank M. L. O'Malley for the computer aided design (CAD) layout work on the NMOS test structure described in Fig. 7; to acknowledge frequent conversations with H. Gossmann and S. Hillenius; and to thank W. Mansfield for facilitating the fabrication effort.

[1]PROPHET is a Bell Laboratories proprietary process simulation CAD tool.
[2]PADRE is a Bell Laboratories proprietary device simulation CAD tool.

[3]B. J. Sapjeta, T. Boone, J. Rosamilia, P. J. Silverman, T. W. Sorsch, G. L. Timp, and B. Weir, "Minimization of Interfacial Microroughness for 13-60A Ultrathin Gate Oxides," Proc. of the MRS, 1997 (to be published).
[4]K. J. Hanson, J. Sapjeta, and G. S. Higashi, in Proceedings of the Symposium on Diagnostic Techniques for Semiconductor Materials, edited by K. K. Schroder, J. L. Benton, and P. Rai-Choudhury (The Electrochemical Society, Pennington, NJ, 1994), pp. 355–369.
[5]Y. Ma and M. L. Green, Proceedings of the 4th International Symposium on Cleaning Technology in Semiconductor Device Manufacturing, edited by R. E. Novak and J. Ruzyllo (The Electrochemical Society, Pennington, NJ, 1996), pp. 115–125.
[6]K. F. Schuegraff, C. C. King, and C. Hu, Technical Digest of 1992 IEEE VLSI Technology Symposium, 18 (1992).
[7]H. S. Momose, M. Ono, T. Yoshitomi, T. Ohguro, C. Nakamura, M. Saito, and H. Iwai, IEEE Trans. Electron Devices 43, 1233 (1996).
[8]D. F. Kyser, J. Vac. Sci. Technol. B 1, 1391 (1983).