Edited by H. Ryssel, P. Pichler - September 1995

Ultra High Performance, Low Power 0.2 μm CMOS Microprocessor Technology and TCAD Requirements

A. Nasr, J. Faricelli, N. Khalil, and C.-L. Huang

Ultra Large Scale Integration Operations Group
Digital Semiconductor
77 Reed Road
Hudson, MA. 01749 U.S.A
(phone: 508-568-5742)

Abstract

The process requirements for low power, ultra high performance CMOS Digital IC's are reviewed. The role, accuracy, and demands of TCAD for future process and device design are discussed. In addition, special challenges for predictive simulation of process and device behavior are highlighted. Two illustrative examples are presented. First, the interaction between the polysilicon gate process architecture and the resulting device characteristics is analyzed in details. Second, subthreshold leakage current (Id_{off}) prediction for devices with 0.2 μm length and below is addressed. A TCAD based inverse modeling strategy is proposed to enhance the predictability of sub-quarter micron CMOS device behavior.

1. Introduction

Ultra high performance, low power CMOS microprocessors are mainstream technology drivers. Operating frequencies up to 1000 Mhz and V_{dd} below 2 V will become a reality by the end of the century. The unprecedented speed of Digital's Alpha microprocessors is due in large part to the following factors: optimized CMOS transistors which balance drive current vs. reliability; process and circuit techniques which consider cost, complexity, yield and reliability; and state-of-the-art TCAD tools that reduce learning cycles.

As the cost of semiconductor fabrication facilities reaches the billion dollar mark, any tool which helps shorten the time between development and product introduction will greatly enhance return on investment. Accurate and dependable TCAD tools play a major role in achieving this goal throughout the product lifetime. During process design and layout rules definition, TCAD tools are used to investigate electrical sensitivity to various conditions. During the reliability optimization phase, simulators are critical for drain engineering and evaluation of hot-carriers reliability. TCAD tools are also used during manufacturing to enhance process robustness to fluctuations due to equipment conditions.

The progress of ULSI technologies to yield higher densities, ultra-high performance and low power chips has brought with it the need to include physical factors previously ignored in TCAD tools. For instance, the reduced channel length and gate oxide thickness have made simple threshold voltage prediction a challenge. This is mainly due to polysilicon gate depletion and dopant penetration from the gate into the silicon substrate. Furthermore, as the threshold voltage is scaled along with the power supply, off-state leakage current Id_{off} prediction and suppression become more critical. Maintaining a high Id_{on}/Id_{off} ratio as the technology and V_{dd} are scaled is of paramount importance.

One reason why the predictive capability of TCAD tools has lagged process development is that our ability to measure what we have made in the fab is limited to large structures with resolution in only one space dimension, or involves time consuming and destructive sample preparation. Without detailed characterization, there is little hope that the models in TCAD tools can be improved. This paper proposes a methodology for improving TCAD tool accuracy and predictive capabilities by a *bootstrapping* technique based on inverse modeling. This allows the extraction of physical and *inaccessible* parameters such as diffusion coefficients in multi-layer films. The same technique is also used to extract the two-dimensional (2D) doping profile which can then be used to predict short channel effects and Id_{off} currents.

2. High Performance and Low Power Requirements

Since the middle of the 1980's, Digital's CMOS technology goal has been characterized by doubling the gate density and improving the gate speed by up to 30% with each successive generation. Fig. 1 shows the maximum operating frequencies and power dissipation as a function of power supply and year of market introduction for a range of high performance Alpha microprocessors and low power portable systems. Fig. 2 shows the normalized saturation current as a function of gate oxide thickness, channel length and power supply.

The interactions between CMOS power dissipation, clock frequency, gate delay and current drive capability are well characterized. Reducing the power supply to lower power consumption will negatively impact the drive current capability and performance. In order to maintain high performance, it is necessary to carefully optimize the gate oxide thickness, threshold voltage and channel length, subject to certain reliability constraints such as hot carrier lifetime and time dependent dielectric breakdown under worse case operating conditions.

The Alpha microprocessor [1] technology uses 3.3 V with a 85 Å gate oxide and 2 V with a 65 Å oxide. Complementary n and p-type implanted amorphous silicon (α-Si) gate interconnect is used for both technologies. Deep ultra violet lithography and anti-reflective coatings are used for good control of channel length down to an L_{eff} of 0.15 μm. The lightly doped and ultra shallow phosphorous or arsenic regions provide both low Id_{off} characteristics and robust hot carrier reliability. BF_2 is used for shallow p-channel device junction and good device characteristics. Rapid thermal annealing is used for the source and drain anneal. A $CoSi_2$ salicide process with sheet resistance below $5\omega/\Box$ is used for low resistance active area and gate interconnect.

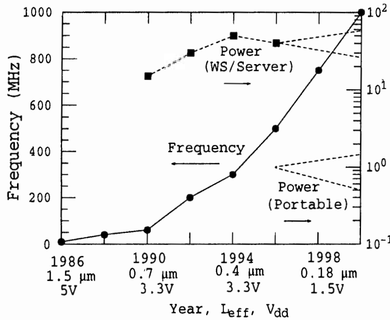

Figure 1: Operating frequency and power dissipation vs power supply and year of introduction.

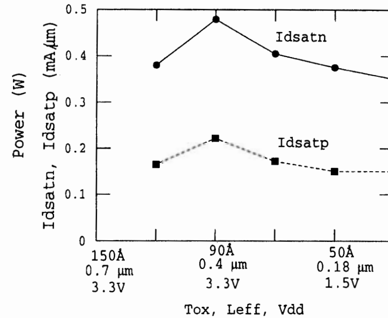

Figure 2: Normalized saturation currents vs t_{ox}, L_{eff}, and V_{dd}.

3. Channel Simulation Issues and TCAD Requirements

3.1. Silicon Gate Interconnect Process Architecture

Thin gate oxides below 80 Å which are used to provide higher current gain at low power supply, are susceptible to dopant penetration during the source and drain implant. The RTA required for the formation of shallow junctions for good short channel effects, good drain-induced barrier lowering characteristics and low Id_{off} gives rise to a non-uniform dopant distribution in the polysilicon gate. The low concentration at the gate/SiO_2 interface accentuates poly depletion effects. This phenomenon results in higher threshold voltage, lower inversion capacitance and reduced current drive capability.

In the process of optimizing 0.2 μm gate length transistor performance for $V_{dd} = 2$ V, the p-channel device design was found to be particularly challenging for TCAD prediction. It was observed that the gate oxide thickness (t_{ox}), silicon gate interconnect deposition temperature and gate film thickness (t_{poly}) were all interrelated in determining the long channel device threshold. Conventional polysilicon gate interconnect, deposited at 620°C, reduced the grain size, as shown in the transmission electron micrograph (TEM) in Fig. 3. Alternatively, silicon gate material deposited at a temperature of 530°C increased the grain size, as shown in the TEM of Fig. 4 where the average grain size is 2000Å. This reduced the boron diffusion and eliminated boron penetration through the gate oxide.

The implanted profile for BF_2 (2×10^{15}, 40 keV) into the polysilicon gate at 530°C and620°C is shown in Fig. 5. Note that while polysilicon grown at 620°C results in more uniform doping distribution, we found that it is more sensitive to penetration and V_{th} shift. On the other hand, the doping profile for the α-Si case shows a lower concentration at the Si/SiO_2 interface. Figure 6 shows the boron doping profile in the 2200Å and 2800Å gate interconnect material. A more uniform doping profile is observed for thinner polysilicon.

3.2. Silicon Gate Interconnect Device Analysis.

Figure 7 shows the impact of polysilicon depletion and boron penetration on the long-channel MOSFET C–V characteristics for devices with 2000Å and 2800Å t_{poly}.

Figure 3: TEM of hi temperature (620°C) polysilicon gate interconnect (Grain size < 300 Å).

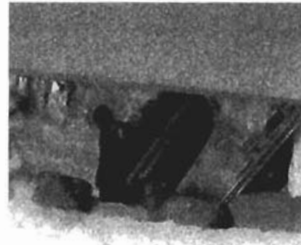

Figure 4: TEM of low temperature (530°C) α-Si gate interconnect (Grain size 2000 – 3000 Å).

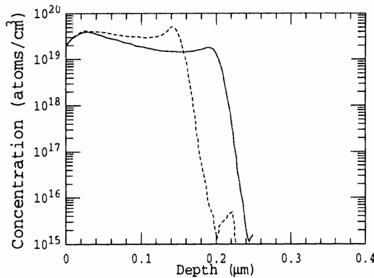

Figure 5: SIMS of BF_2 implant with 2000 Å (dashed) and 2800 Å (solid) α-Si.

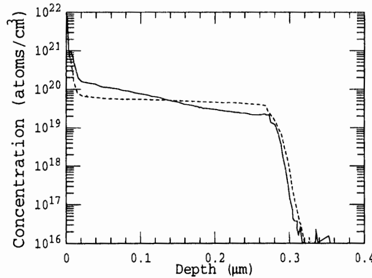

Figure 6: SIMS BF_2 comparison in α-Si (dashed) and polysilicon interconnect (solid).

The processing conditions and extracted parameter values for both Devices A (solid line) and B (dashed line) are given in Table I. Note that as t_{poly} is reduced, the polysilicon concentration, N_p, increases. The increase in N_p results in substantial improvement in the inversion capacitance for Device A (see Fig. 7). The major drawback of thinning t_{poly}, however, is the penetration of implanted specie (boron in this case) from the gate to the channel region, thus affecting the threshold voltage stability [2].

Since t_{poly} has such dramatic impact on the polysilicon depletion and boron penetration for the p-channel MOSFET with a thin gate oxide, more careful study of the source/drain implant condition is in order. Figure 8 shows comparison of MOSFET C–V degradation for three implant conditions. The process conditions for these devices are also listed in Table I as Device C (long-short dashed line), D (solid line) and E (short dashed line). It is important to note that by changing implant condition from Devices C to E, the degradation of C–V characteristics in the inversion regime increases approximately 5%. In addition, we note that, despite the increase in the N_p concentration for Device C, the boron penetration is still invisible in Fig. 8. This, however, is not the case for Device A.

We notice that by changing the t_{poly} from 2200Å (Device C) to 2000Å (Device A), both N_p and boron penetration increase. In other words, the sensitivity of

Table I: Comparison of extracted parameter values for Devices A to E.

Devices	A	B	C	D	E
t_{ox} (Å)	80.2	80.2	83.4	83.3	83.4
t_{poly} (Å)	2000	2800	2200	2200	2200
P^+ S/D Dose	2.5×10^{15}	2.5×10^{15}	2.5×10^{15}	2.5×10^{15}	2×10^{15}
P^+ S/D Energy	40	40	40	35	35
N_p ($10^{19}\times$ cm^{-3})	2.4	1.1	1.4	1	0.85

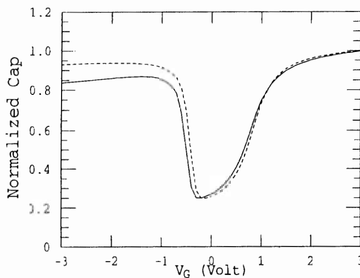

Figure 7: Gate capacitance as a function of polygate thickness (solid line – Device A and dashed line – Device B).

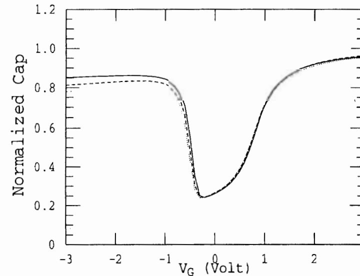

Figure 8: Gate capacitance as a function of S/D implant conditions (long-short dashed line – Device C, solid line – Device D and dashed line – Device E).

polysilicon depletion and boron penetration can be affected by a mere 200 Å t_{poly} difference. It, therefore, presents a tremendous challenge in designing p-channel MOSFET with conventional gate oxide that have sufficiently high N_p and without boron penetration.

3.3. P-Channel Inverse Modeling

One-dimensional simulation using SUPREM-III [4] of the above mentioned process conditions cannot consistently reproduce the observed electrical device behavior. On the other hand, an inverse modeling characterization methodology was successful in determining the profiles as explained in the following.

The average gate polysilicon doping concentration was first determined by matching simulated and experimental capacitance data in the inversion region. Thereafter, the one-dimensional doping profiles in the channel region were extracted using inverse modeling from deep depletion capacitance data [5]. In Fig. 9, the two extracted profiles corresponding to 2000Å and 2800Å t_{poly} are shown. The effect of boron penetration is apparent as a reduction in the net doping concentration near the Si/SiO_2 interface.

Using the extracted profiles, and the average polysilicon doping, the quasi-static C–V characteristics were simulated. By comparing the simulated and experimental measurements, it was found that for the device with boron penetration, the simulation cannot reproduce the experimental threshold voltage shift. Indeed, as a result

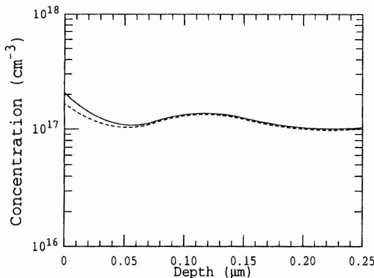

Figure 9: Comparison of channel doping profile for device with 2800Å (solid) and 2000Å (dashed) t_{poly}.

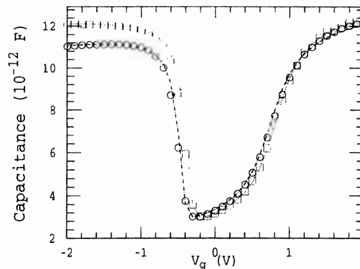

Figure 10: simulated (solid) and experimental (symbols) quasi-static C–V device with 2800Å (solid) and 2000Å (dashed).

of low, non-degenerate doping, modeling the resultant nonuniform polysilicon doping is essential for accurate device simulation. By the same token, one can use the experimental C–V data to determine the polysilicon doping profile.

The spline profile parameterization [5] was extended to represent the polysilicon doping profile. The values of the B-splines coefficients can then be extracted from experimental quasi-static data. Using this approach the polysilicon doping is determined, and an excellent fit of the experimental quasi-static C–V data is achieved as shown in Fig. 10.

It is noted that the procedure described above to determine polysilicon and channel doping profiles for various processing conditions can provide the required data for *tuning* SUPREM-III diffusion coefficients.

4. Subthreshold Simulation and TCAD Requirement

4.1. 2D physical characterization

Two of the most important quantities that must be characterized for accurate simulation of a deep submicron MOSFET are the polysilicon gate length and the dopant profile. In this discussion, we focus on 2D effects, and ignore 3D effects such as the narrow width effect on threshold voltage. In any process, but especially one under development, there can be considerable variation from the "as-drawn" poly gate length to any given poly gate on the wafer. When analyzing the characteristics of any particular transistor from a distribution of MOSFETs, it is imperative to know the actual gate length. In earlier process development efforts, we used techniques such as cross-sectional SEM to quantify the gate length. This technique, while accurate, is time consuming and can only be performed on limited numbers of devices. It also results in destroying the device under test. An alternate technique we have devised uses the measurement of the device gate capacitance in inversion to infer the gate length [6]. We first extract the gate oxide thickness t_{ox} from capacitance-voltage (C–V) measurements on a large area MOS capacitor. While extracting t_{ox}, we also extract the average polysilicon doping near the oxide interface. We then use the MINIMOS [7] device simulator to simulate the inversion capacitance of the

short channel device. We vary the simulated gate length until a match is achieved with the measured capacitance. We take great lengths to include effects which would alter the inversion capacitance, such as polysilicon depletion, quantum mechanical effects important for thin t_{ox}, and various fringing capacitances [8]. This technique has been verified using cross-sectional SEM, and has shown to be effective in extracting gate lengths down to 0.1 μm.

In the area of dopant profiling, we have had considerable success in using the inverse modeling technique to extract the 2D doping profile [9]. The technique is described in the reference, but a brief description is given here for completeness. The poly gate length extraction technique described above provides the t_{ox} and polysilicon dopant concentration for the device. The channel dopant for a long channel device is extracted from deep depletion capacitance measurements. In the source/drain regions far from the gate, SIMS measurements of the source/drain dopants are used along with the area capacitance of the source/drain diode to extract the complete profile in that device region. These profiles are combined with the channel dopant profile to form an initial guess at the 2D profile. Capacitance measurements which "probe" various parts of the profile, such as gate capacitance vs. gate voltage in accumulation, are used as inputs to the 2D extraction. A device simulator (in our case, MINIMOS) is used to simulate these capacitances. A non-linear optimizer is used to determine the dopant profile which produces simulated capacitances that best fit the measured ones. A key feature of the technique is the tensor product splines representation of the 2D dopant profile which allows a completely general form for the dopant distribution, while reducing the description of the profile to a handful of coefficients. This greatly reduces the amount of work the optimizer must do compared to a straightforward mesh representation.

The results of such an extraction are shown in Fig. 11. We note that the dopant extracted in the source/drain extensions under the gate are relatively immune to small changes in the measured data or other conditions in the extraction procedure. The extracted channel dopant, on the other hand, is sensitive to assumptions about the amount of fixed charge at the Si/SiO_2 interface, or the value of the polysilicon workfunction. One strategy we have recently begun to use is to assume a small fixed interface charge and use a polysilicon workfunction that includes bandgap narrowing effects. We then let the channel dopant adjust itself to fit the measured threshold.

4.2. Id_{off} simulation issues

One important design criteria of n-channel MOSFET is the "off" current Id_{off}. Maintaining a low Id_{off} becomes increasingly difficult as device thresholds are reduced below 0.5V. Careful design of the inner source/drain junction is required to reduce short channel effects. We use two-dimensional process simulation to help design the inner junction, but many physical effects, such as boron depletion from the channel due to adjacent source/drain dopants [10], are not yet included in commercial simulators. Using the inverse modeling techniques described in the previous section, we can get some idea of what the source/drain and channel profiles look like. We can then feed the results into a device simulator and see if the correct threshold and Id_{off} behavior is predicted. This further helps validate our extraction procedure. Thus, even if we cannot yet do predictive process simulation from first principles, we can at least understand the impact of process conditions on device characteristics.

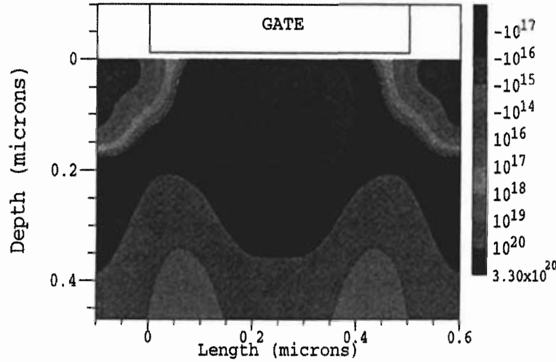

Figure 11: 2D extracted profile.

In Fig. 12, we show a plot of n-channel V_{th} vs. L_{eff} for a phosphorous moderately doped drain (MDD) structure, using a 7 degree implant in our 0.35 μm process. The figure also shows the simulated threshold voltage using a two dimensional doping profile extracted from capacitance data described in the previous section. The extraction was done on a test structure which had a gate length of 0.47 μm. We then simulated smaller channel lengths by simply scaling the 2D profile, as we had done for earlier technologies. As can be seen from the figure, scaling the profile breaks down at channel lengths below L_{eff} of 0.3 μm.

In Fig. 13, we show a plot of measured n-channel Id_{off} vs. L_{eff}, where Id_{off} is defined as the drain current at V_{gs}=0V, V_{ds}=2.7V, and 100° C. The figure also shows the simulated Id_{off} (solid line). The simulated Id_{off} is about 2.5–3 times higher than measurement down to L_{eff}=0.3 μm. This is reasonable, considering that the Id_{off} is very sensitive to the details of the doping profile. The simulated Id_{off} at L_{eff}=0.22 μm, however, is several orders of magnitude larger than measurement. This is expected, since the simulated threshold voltage is 120mV lower than measurement. An interesting observation is that not all of the increase in Id_{off} can be attributed to the inaccuracy in the threshold. Indeed, simulated Id_{off} at L_{eff}=0.22 μm with the threshold adjusted to match the measured data is still over an order of magnitude larger than measurement.

One can conclude that the doping profile changes considerably when the device length varies only a few tenths of microns. Thus, our initial methodology of scaling a profile extracted from one gate length is not appropriate for deep submicron MOSFET's. A better methodology is to use test structures that span the range of expected channel lengths from "short" to "long" and extract profiles for each channel length. We are in the process of designing a test chip with such features. This technique can be used to improve process simulators physical models which will then allow better prediction of subthreshold characteristics.

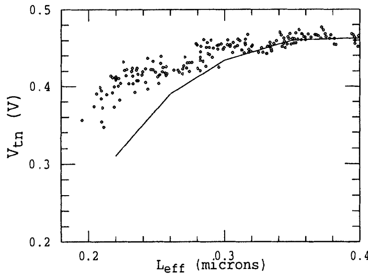

Figure 12: V_{th} v/s L_{eff} with 65 Å t_{ox}.

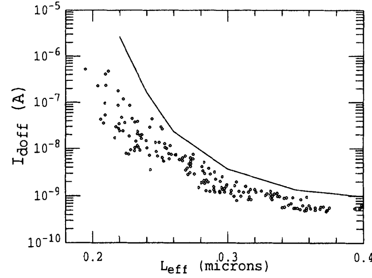

Figure 13: Id_{off} v/s L_{eff} with 65 Å t_{ox} (V_{gs}=0V, V_{ds}=2.7V, and 100° C).

5. Conclusion

We have outlined a strategy to improve TCAD prediction for sub 0.2 μm devices. Examples of polysilicon depletion, boron penetration and Id_{off} prediction requirement were addressed in details. Inverse modeling coupling with TCAD tools was proposed as a means to calibrate physical process parameters to device electrical characteristics.

References

[1] D. Dobberpuhl *et al.*, *IEEE Int'l Solid State Circuits Conf.* p. 106, 1992.
[2] J.R. Pfiester *et al.*, *IEEE Trans. Electron Devices* Vol. 37, p. 1842, 1990.
[3] G.Q. Lo *et al.*, *IEEE Elect. Dev. Letters* vol. 12, p. 175, 1991.
[4] C.P. Ho and S.E. Hansen, *SUPREM-III, Tech. report 83-001*, Stanford, 1983.
[5] N. Khalil *et al.*, *IEEE Elect. Dev. Letters*, Vol. 16, p. 17-19, 1995.
[6] C.-L. Huang *et al.*, submitted to *IEEE Trans. Elect. Dev.*, 1995.
[7] S. Selberherr, *et al.*, *IEEE Trans. on Elect. Dev.*, vol. 27, p. 1540, 1980.
[8] R. Rios and N. Arora, *IEDM Techn. Digest*, p.613-616, 1994.
[9] N. Khalil *et al.*, *Accepted for presentation at ESSDERC'95.*
[10] H. Hanafi *et al.*, *IEEE Elect. Dev. Letters*, Vol 14, No. 12, p.575-577, 1993.