Simulating the Hysteresis effects of Si/SiO_2 Interface Traps

Introduction

The trap states at the interface of Silicon with Silicon Dioxide are usually divided conceptually into interface states and fixed oxide charges [1]. These trap states can cause the degradation of the performance of devices such as MOSFETs, when they are stressed into a regime where hot carriers are significant. This degradation is usually permanent and occurs for stress times of the order of 10^3 seconds [2]. This phenomenon is well studied, and ATLAS has a degradation model for simulating these effects [3].

If a MOSFET is stressed so as to avoid significant hot carrier current, the threshold voltage V_t can show hysteresis effects on the sub-second time scale [4]. This can affect the settling time of circuits containing MOSFETs. It is proposed in [4] that the mechanism for the hysteresis is tunnelling of electrons directly from the channel into the fixed oxide traps, as well as indirectly, via the interface states. The experimentally observed dependence of hysteresis on measurement time is apparently predicted by the model of Heiman et al [5]. This model is implemented in ATLAS and can be used to simulate hysteresis effects due to oxide trapping. [6] In the next section we describe this model and its implementation in ATLAS. Some examples of its use are then described.

Model for Hysteresis Behavior of Traps

The equation for the transient behavior of acceptor-like interface traps is:

$$\frac{dF_{tA}}{dt} = v_n SIGN\left[n(1-F_{tA}) - DEGEN.FAC F_{tA} n_i exp(\frac{E_{tA}-E_i}{kT})\right]$$
$$- v_p SIGP\left[pF_{tA} - \frac{(1-F_{tA})n_i}{DEGEN.FAC} exp(\frac{E_i - E_{tA}}{kT})\right]$$

where F_{tA} is the occupation probability, v_n and v_p are the thermal velocities for electrons and holes respectively and n and p are the electron and hole concentrations respectively. DEGEN.FAC is a degeneracy factor and E_i and E_{tA} are the intrinsic and trap levels respectively. The trap cross-sections for capture of electrons and holes are SIGN and SIGP respectively. The acceptor-like traps are electrically neutral when empty, and negatively charged when occupied by an electron. There is an equivalent expression for donor-like traps, although in this case we only consider acceptor-like traps[6].

The four distinct terms on the right hand side of this equation result in four different rates; a charging rate by conduction band electrons, a discharge to the conduction band, a discharge rate due to valence band hole capture, and a charging rate due to emission of holes to the valence band.

These rates depend linearly on the cross-sections SIGN and SIGP, and which rate is dominant depends on the carrier surface densitites, as well as the parameter values. Asymmetry of charging and discharging rates leads to hysteresis effects of fast interface states, but on a timescale much shorter than that considered here [4]. In order to get experimentally observed behaviour it is necessary to consider the interface states extending spatially into the oxide, with the values of SIGN and SIGP modified by the tunnelling mechanism needed for the carriers in the channel to interact with the buried traps. In the HEIMAN model, the values of SIGN and SIGP are assumed to depend on their distance into the insulator, d, as

$$SIGN(d) = SIGN \quad e^{-\kappa_e d}$$
$$SIGP(d) = SIGP \quad e^{-\kappa_h d}$$

where

$$\kappa_e^2 = \frac{2m_e^*(E_c - E_{tA})}{\hbar^2}$$
$$\kappa_h^2 = \frac{2m_h^*(E_{tA} - E_v)}{\hbar^2}$$

The tunnelling to the trap states from the conduction band edge in the channel E_c is modelled by the term with the evanescent wavector κ_e for electrons. Tunnelling from the valence band edge E_v to the trap E_{tA} is modeled by the evanescent wavevector, κ_h for holes. The trap spatial density is assumed to be uniform up to a certain depth, and zero above that, and the traps are assumed to be monoenergetic. The capture cross-sections fall off exponentially with distance, and thus to significantly change the occupation state of a trap at depth d requires a simulation time depending exponentially on d. This dependence of the capture-cross sections on depth leads to hysteresis effects on much longer timescales than can be accounted for by the surface states. Differences between κ_e and κ_h also increase differences between charging and discharge rates.

Examples

The first example is of a MOS Capacitor with acceptor-like interface traps.

It has a 20 nm thick oxide layer and the silicon region is uniformly doped with 1 x $10^{17} cm^{-3}$ acceptors. The interface traps are included in the ATLAS input deck with the INTTRAP statement

```
INTTRAP E.LEVEL=0.6 ACCEPTOR
DENSITY=1.0e12 S.I DEGEN=4 SIGN=1.0e-13
SIGP=1.0e-14
```

which defines interface traps at a density of 1 x $10^{12} cm^{-2}$, 0.6 eV below the Silicon Conduction band edge. The capture

cross sections, SIGN and SIGP are also specified in units of cm^2. The gate is biased to 2 V in steady-state, which ensures that the traps are initially filled with electrons. The gate bias is ramped to -2 V, and then immediately back to 2 V. It is also possible to simulate a settlement time after the first ramp, but this is not considered here. Without the Heiman model enabled there is no hysteresis, and the gate-substrate capacitance obtained from small-signal analysis at 100 Hz is shown as a function of gate bias in Figure 1a. The capacitance of the MOS-C without interface traps is shown for comparison. The presence of the traps contributes significantly to the Capacitance over the Voltage range where the trap occupation is changing, with a maximum when the Fermi-level coincides with the trap energy. If we now enable the Heiman model, by adding the parameters

```
HEIMAN DEPTH=0.001 HPOINTS=10
```

to the INTTRAPS statement, we use the SOLVE statement to carry out the bias ramps with the ramptime set to 0.1 seconds by using the TIMESPAN parameter.

```
SOLVE VSTEP=-0.05 NAME=GATE VFINAL=-2.0
AC.ANAL FREQ=1.0e2 DIRECT TIMESPAN=0.1
```

We then observe hysteresis effects. The HEIMAN model spreads the interface trap density over the specified depth, using the number of depth slices specified. In this case the uniform distribution is to a depth of 1nm, with 10 slices of 0.1 nm each. In Figure 1b the average trap occupancy is shown as a function of bias. On the sweep from 2V to -2V (red curve), the average occupancy decreases more slowly than the non-hysteresis case (blue curve). At -2V the deepest traps are not completely emptied and so the average occupancy is still finite. For the first part of the -2V to 2V sweep (green curve) the average occupancy is still decreasing because the Fermi-level is still below the trap level. The average occupancy then increases, lagging the case of no hysteresis because the deeper traps are slower to fill. The effect of hysteresis on the small signal capacitance is shown in Figure 1c. The capacitance is reduced because of the increased lifetimes of the deeper traps, and there is a voltage shift of approximately 0.5 V between the results for the two ramp directions. The surface Voltage also shows hysteresis effects as shown in Figure 1d.

The second example is a PMOSFET with a channel length of 0.25 microns. This is the same structure as in example 2 of the ATLAS Advanced MOSFET examples, but with acceptor-like interface traps placed along the channel part of the interface. The traps have the same parameters as in the above example, except that the density is reduced to $10^{11} cm^{-2}$ and

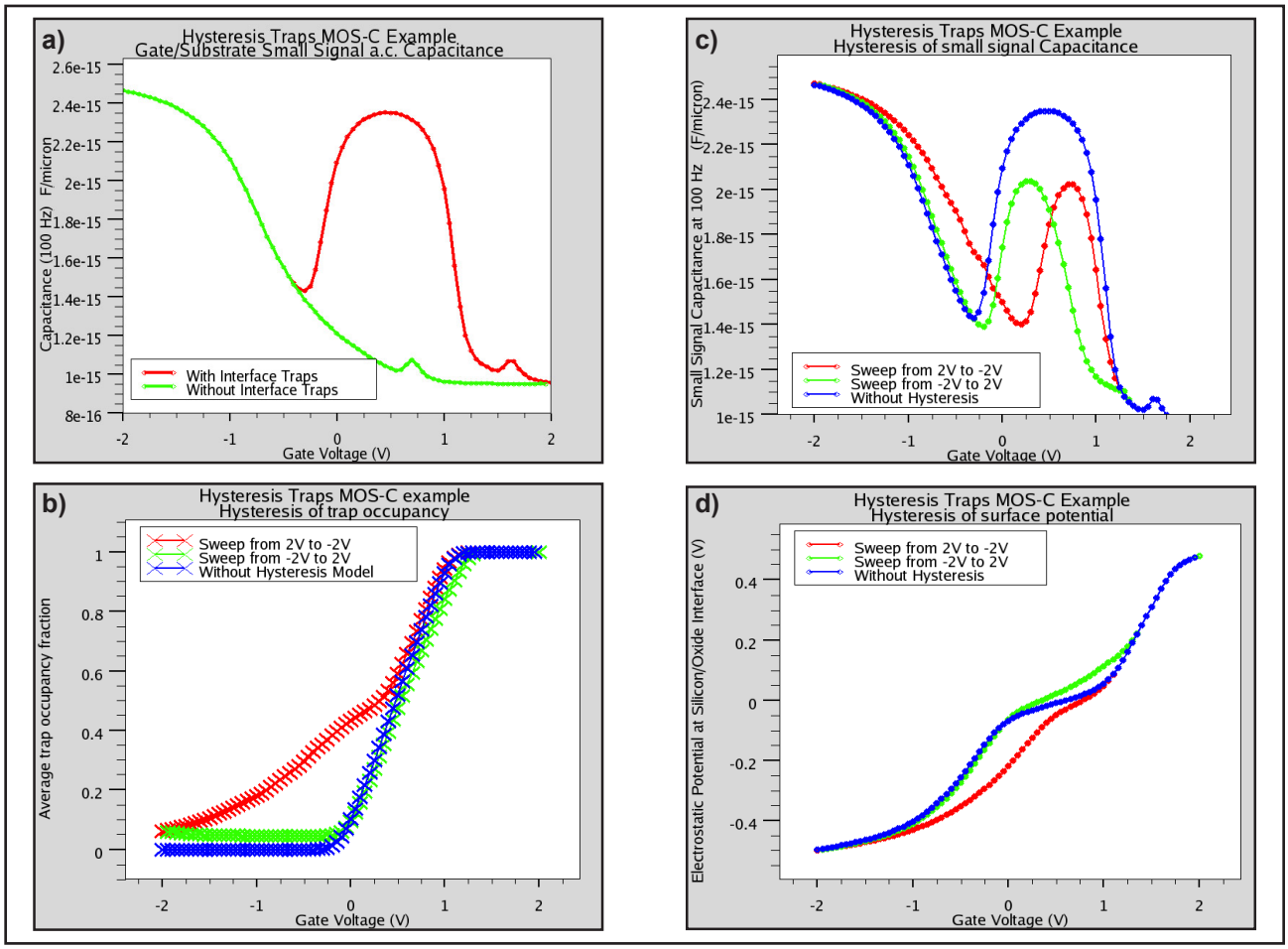

Figure 1: Results for the MOS-C example.

that the traps were assumed to extend a distance of 2 nanometers into the interface. The source-drain bias was set to 0.5 Volts quasistatically and then the gate voltage was ramped to turn on the device. The source-drain bias was kept low in order to avoid the regime where hot electrons are important, and so that the drift-diffusion approximation was valid. The threshold voltage was defined as the current required to achieve a source-drain current of 10 microamperes per vertical micron. The gate bias was linearly ramped to 3 V in a set time (ramptime), and V_t obtained from the curve of drain current versus gate bias. The device was allowed to settle in the on-state for a long enough time to attain a steady-state. The gate bias was then linearly de-ramped back to zero Volts in the same set time, and V_t obtained from the curve. The sequence of ATLAS SOLVE statements is

```
SOLVE VGATE=0.0 VSTEP=0.1  NAME=GATE
VFINAL=3.0 TIMESPAN=$RAMPTIME

SOLVE PREV

SOLVE VGATE=3.0 VSTEP=-0.1 NAME=GATE
VFINAL=0.0 TIMESPAN=$RAMPTIME
```

The V_t obtained for the ramp, minus the V_t for the deramp, is shown in Figure 2a as a function of ramptime. This hysteresis in the V_t can be explained as follows. The interface traps are initially empty of electrons, but as the gate bias is ramped they fill with electrons and so shift the V_t to a higher value. The charging rate depends linearly on the density of channel electrons as well as the the capture cross-section. The capture cross-section decreases rapidly with depth, and for shorter ramptimes the deeper traps are not filled with electrons. This reduces the interface charge and thus reduces the V_t. After the interface charge is allowed to attain its steady state value, the gate bias is deramped. The discharging is mainly by electron emission to the conduction band because the hole density is too small for significant hole capture. The traps do not discharge significantly until the gate bias is below threshold, and so there is no ramptime dependence of V_t. This is illustrated in Figure 2b, where we see the average occupancy of the interface trap halfway between the source and drain. The ramptime is 1 millisecond and the quasistatic solution is shown for comparison. The ramp from 0V to 3V on the millisecond timescale shows that the trap occupancy significantly lags the quasistatic case. The deramp from 3V to 0V shows that the trap occupancy is the same for the quasistatic and 1 millisecond cases above V_t (which is around

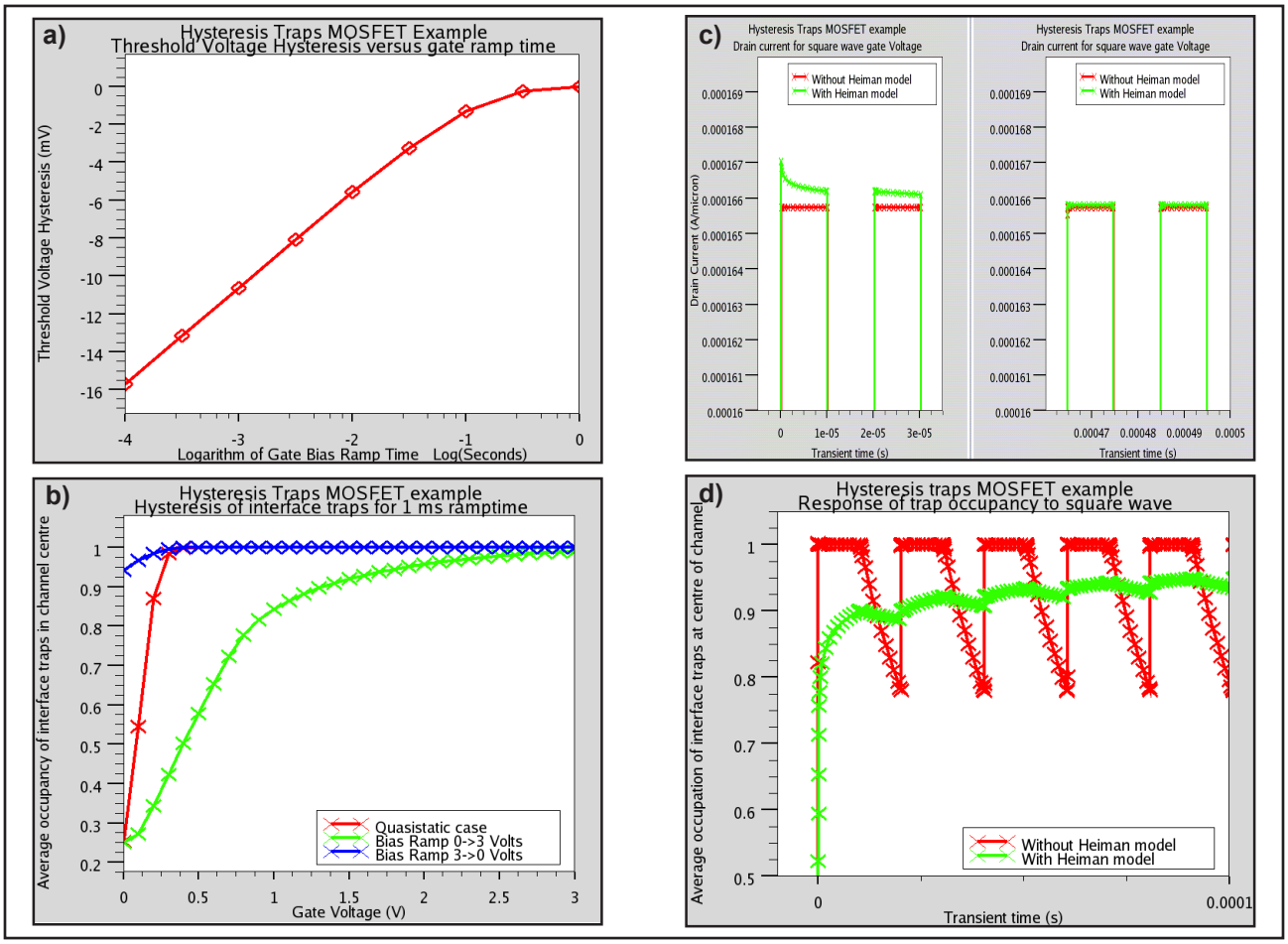

Figure 2 : Results for the MOSFET example.

1V). Only at lower biases does the quasistatic curve show a more rapid discharging, due to the capture of holes as the hole density at the interface increases. Figure 2a shows a dependence of hysteresis shift on the logarithm of ramptime over a range of ramptimes, as also reported by [4]. For longer ramptimes, the V_t shift tends asymptotically to zero.

The HEIMAN model is also available for fully transient solutions, which permits the calculation of large signal parameters. A square wave of amplitude 3V was applied to the gate of the MOSFET. The rise and fall times were 0.1 μs and the voltage was held at 3V or 0V for 10 μs periods. We compare the resulting drain currents for the first two square waves with two square waves occurring after about 500μs in Figure 2c. Two things are immediately apparent. Firstly, the first pulse has some overshoot of current with a rapid decay to a constant value. Secondly, after the simulated time has reached 500μs, the effect of hysteresis has almost vanished. This can be understood by looking at the average occupancy of the interface traps, as shown in Figure 2d. Without hysteresis the trap occupancy rises more quickly to its maximum value, and the threshold voltage is increased because of the larger negative charge trapped at the interface. With hysteresis the charge trapping is delayed, and the lower threshold voltage gives rise to an initially larger drain current. The trap occupancies without hysteresis effects are modulated as the channel moves into and out of inversion, as the simulation progresses. This is because the channel hole density is sufficient to cause the traps to discharge at zero gate bias, as seen in Figure 2b. For the hysteresis traps this modulation is much less pronounced, and the electron density at the interface slowly increases to the average occupancy. Thus, after 500us, the drain currents are approximately the same in both cases as seen in Figure 2c.

Conclusions

ATLAS is capable of simulating the hysteresis effects of interface traps. The mechanism of the hysteresis is tunnelling to and from traps located inside the oxide. This results in a spread of charging times as well as greater asymmetry between charging and discharging times. These effects differ from hot carrier degradation effects because they are not permanent, and occur on a shorter timescale. The effect on small signal ac parameters can be modelled, and fully transient large signal analysis can also be carried out. For an example MOSFET, the hysteresis in the threshold voltage has been shown to depend on the logarithm of the measurement time over a range. This agrees qualitatively with experimental observations [4].

References

[1] C. R. M. Grovenor, 'Microelectronic Materials' , Adam Hilger, (1989), pp. 300-301.

[2] W. Hansch, A. v. Schwerin and F. Hofmann, 'A new Self-Consistent Modeling Approach to Investigating MOSFET Degradation', IEEE Electron Device Letters, Vol. 11, No. 9, pp. 362-364, (1990).

[3] *ATLAS Users Manual*, Section 3.6.7, Silvaco Inc (2010).

[4] T. L. Tewksbury, Hae-Seung Lee and G.A. Miller,'The effects of oxide traps on the large-signal transient response of analog MOS circuits', IEEE Journal of Solid-State Circuits, Vol. 24 No. 2, pp. 542-544, (1989).

[5] F.P. Heiman and G. Warfield, "The effects of Oxide Traps on the MOS Capacitance", IEEE Trans. Electron Devices, Vol. 12, No. 4, pp. 167-178, (1965).

[6] *ATLAS Users Manual*, Section 3.3.3, Silvaco Inc (2010).