ACCOUNTING FOR INCOMPLETE IONIZATION IN MODELING SILICON BASED SEMICONDUCTOR DEVICES

Daniel C. Cole and Jeffrey B. Johnson
IBM General Technology Division, Essex Junction, Vermont 05465

ABSTRACT

Here, we will first briefly review the approaches researchers have used to quantitatively describe incomplete ionization. We'll then describe a fairly simple approach that can easily be incorporated into device modeling programs, and that appears to yield fairly good results. At low doping concentrations, the model is equivalent to the incomplete ionization formulae frequently presented in device engineering textbooks. At high doping concentrations, the model correctly predicts that all impurities are ionized, as is observed experimentally. The model is a phenomenological one, with two parameters that make connections with physical parameters at low and high doping concentrations. We've found a least squares method that yields these parameters. Finally, we'll show some device simulation results and describe how incomplete ionization must enter into the current continuity equations for electrons and holes.

INTRODUCTION

The incomplete ionization of a dopant impurity in a semiconductor is well understood physically for the case where the impurity concentration is low. In this case, the impurity states are localized, their energy level is discrete, and the band structure for the semiconductor is essentially unaltered from the intrinsic case. Counting the total number of ways of arranging the electrons in the available energy levels is then fairly simple. Maximizing this quantity, subject of course to constraints like the total number of electrons present and the total energy available, then yields the ionized fraction of the impurity concentration. In [1], Look gives an excellent description of this procedure: the general case of multicharge centers with excited energy states is thoroughly discussed here, and both localized and band states are handled via a uniform treatment.

The following formulae are used in such standard references as [2] and [3] to describe the fraction of ionized shallow donor and acceptor substitutional impurities in silicon:

$$N_D^+ = \frac{N_D}{1 + F_D \exp\left[\dfrac{\Delta E_D + (E_{fn} - E_C)}{kT}\right]}, \quad (1) \qquad N_A^+ = \frac{N_A}{1 + F_A \exp\left[\dfrac{\Delta E_A + (E_V - E_{fp})}{kT}\right]}. \quad (2)$$

Here, (1) $\Delta E_D \equiv (E_C - E_D)$ and $\Delta E_A \equiv (E_A - E_V)$ are the donor and acceptor impurity ionization energies, (2) E_C and E_V are the conduction and valence band edge values, and (3) E_{fn} and E_{fp} are the electron and hole quasi-fermi energy levels, respectively. Due to the fact that a much larger energy is needed to place two electrons in a localized state, as opposed to just one electron, then $F_D = 2$, as can be seen from Eqs. (17) and (21) in [1]. For F_A, the argument is often given that $F_A = 4$ in silicon, since there is a heavy and light hole degeneracy of the valence band, with two possible spin states in each [3].

Equations (1) and (2) fit measured ionized impurity concentrations reasonably well for donor and acceptor impurity concentrations less than or equal to about 10^{17} cm^{-3}. Also, these expressions can be improved slightly by taking into account excited energy states for the impurities, as discussed in [1]. However, at concentrations around 10^{18} cm^{-3}, several effects become important that greatly complicate the situation. Neighboring impurities in the semiconductor are now close enough that electron wave functions at the impurity states overlap, resulting in their initial discrete energy levels changing into energy bands. Moreover, as a result of the random distribution of impurities, fluctuations occur in the electrostatic potential throughout the crystal. Consequently, we must expect some spatial fluctuation in value of the conduction, valence, and impurity bands, thereby giving rise to what is often referred to as bandtails. In addition, interactions between carriers and between ionized impurities and carriers results in shifts of the conduction, valence, and impurity bands. Indeed, for large enough doping concentrations, the impurity band enters into the majority carrier band, so the original concept of separate bands loses meaning.

To accurately describe this phenomena is a very challenging problem to solid-state physicists. References [4], [5], and [6] contain theoretical studies related to this subject. However, for device modeling programs, such intensive computations are probably too involved to be of significant aid. Indeed, these computations do not even include the problem of electron transport.

In order to incorporate incomplete ionization at high doping concentrations into the realm of device modeling applications, Kuzmicz [7] took a more pragmatic approach to the problem. He combined several models in the literature and made a few simplifying assumptions, so as to obtain a more phenomenological description of incomplete ionization. His approach contains only one adjustable parameter.

Here, we describe another approach to the problem that is considerably simpler than even Kuzmicz's method. Its disadvantage is that specific pieces of the contributing factors to incomplete ionization are not separately modeled. Thus, for example, the model does not attempt to describe the broadening of the impurity energy bands, or the tails that form in the conduction and valence band. However, the model does make a natural connection with the important physical parameters at low and high doping concentrations. Moreover, the model gives a fairly good description of the important qualitative features of the incompletely ionized impurities, while still remaining simple and easy to incorporate into present drift-diffusion silicon based device modeling programs, such as PISCES, FIELDAY, HFIELDS, MINIMOS, etc.

INCOMPLETE IONIZATION MODEL

In order to retain the simplicity of Eqs. (1) and (2), but to also model incomplete ionization effects at higher concentrations, we used the phenomenological approach of treating F_D, F_A, ΔE_D, and ΔE_A as being dependent on impurity concentrations. For donor concentrations near $10^{19}\,cm^{-3}$, then $F_D \rightarrow 0$; likewise, $F_A \rightarrow 0$ for acceptor concentrations near $10^{19}\,cm^{-3}$. In this way, the following experimentally observed result at high impurity concentrations is easily modeled: namely, that all of the impurities are ionized, even at low temperatures. Thus, Eqs. (1) and (2) agree with theory at high concentrations, where the impurity bands merge with the majority bands to yield complete ionization, as well as at low concentrations. Between these domains, the models must be simply viewed as empirical expressions that were fit to data.

To obtain the empirical parameters as functions of doping concentration, a simple least squares method was used to help fit Hall data [8], [9] over the temperature range of $50°K \rightarrow 300°K$. For uniformly doped samples of silicon with a very small potential bias applied, the electric displacement $\overline{D}$ within the sample will be essentially constant. From Gauss' law, $\nabla \cdot \overline{D} = \rho$, so that the charge neutrality condition applies of $\rho = q(p - n + N_D^+ - N_A^-) \simeq 0$. Using the effective mass approximation of rigid, parabolic bands [10] to describe electron and hole carrier concentrations, then

$$n = N_C F_{0.5}\left(\frac{E_{fn} - E_C}{kT}\right), \qquad (3) \qquad\qquad p = N_V F_{0.5}\left(\frac{E_V - E_{fp}}{kT}\right), \qquad (4)$$

where N_C and N_V are the effective density of states for holes and electrons, and $F_{0.5}$ is the Fermi integral of order 1/2. Under low transport conditions, E_{fn} and E_{fp} are essentially equal. We then obtain the equation below that can be solved for ξ, where $\xi = E_f - E_V$:

$$0 = N_V F_{0.5}\left(\frac{-\xi}{kT}\right) - N_C F_{0.5}\left(\frac{\xi - E_g}{kT}\right) + \frac{N_D}{1 + F_D e^{\frac{\Delta E_D}{kT}} e^{\frac{\xi - E_g}{kT}}} - \frac{N_A}{1 + F_A e^{\frac{\Delta E_A}{kT}} e^{\frac{-\xi}{kT}}}, \qquad (5)$$

If p doped samples are being analyzed, with $N_A \gg N_D$, then the quasi-Fermi energy level E_f will be near the valence band energy level. In the third term in Eq. (5), $\Delta E_D + \xi - E_g = -(E_D - E_f)$. We can approximate this quantity by replacing E_f by E_V. For shallow energy donor states, then $E_D \simeq E_C$. Hence, $\Delta E_D + \xi - E_g \simeq -(E_C - E_V) = -E_g$. Since $E_g/kT \gg 1$, at least for the temperature range being considered here where T is less than $300°K$, then

$$\exp\left[\frac{\Delta E_D + \xi - E_g}{kT}\right] \simeq \exp\left[\frac{-E_g}{kT}\right] \ll 1. \tag{6}$$

Hence, the third term in Eq. (5) can be replaced by N_D. Solving for ξ in the last term in Eq. (5) and replacing the first two terms by p and $-n$, respectively, then yields

$$\left\{\frac{\xi}{kT} + \ln\left[\frac{(N_A - N_D) - (p-n)}{(p-n) + N_D}\right]\right\} = \left(\frac{1}{kT}\right)\Delta E_A + \ln F_A. \tag{7}$$

In order to fit ΔE_A and F_A to a set of hole concentration data, the following procedure can be followed. N_A and N_D must be known, p must be experimentally determined, ξ can then be obtained from Eq. (4), and n can then be obtained from Eq. (3), or it can be ignored, as it's much smaller than p. Hence, the LHS of Eq. (7) can be evaluated. By plotting this quantity versus $1/kT$, then a least squares fit to data yields ΔE_A as the slope, and $\ln F_A$ as the y-intercept. An example of this method is shown in Figs. 1 and 2 for $N_A = 10^{18}$ cm^{-3}, This value is illustrated here since it lies at the start of the high doping regime, but low enough that the impurity band has not simply merged completely with the valence band, and high enough so that a discrete acceptor energy state is not valid. In Fig. 1, Y_a equals the LHS of Eq. (7). The data was taken over the range of $50 \rightarrow 300°K$ from Fig. 7, sample #141, in Ref. [9.] This plot yields that $F_A = e^{y\text{intercept}} = e^{-0.75} = 0.47$, and $\Delta E_A = .039$ ev . Using these values for ΔE_A and F_A, Eq. (5) was then solved to obtain the hole concentration p, which is shown as the continuous curve in Fig. 2. The data in Fig. 2 again represents values read off Fig. 7 for sample #141 in [9]. As can be seen, the fit obtained in Fig. 2 between data and model is good. A similar procedure can be followed when analyzing electron concentration data.

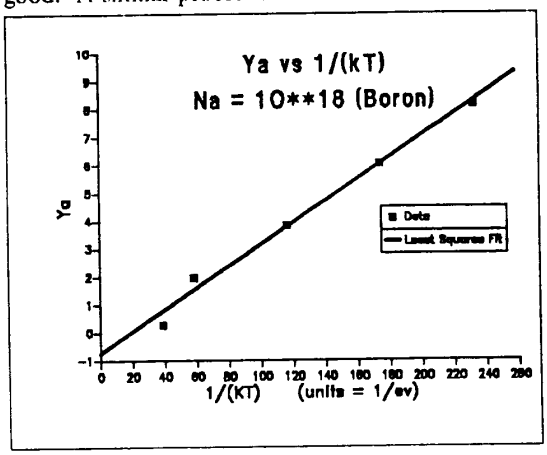

Figure 1. Extracting ionization model parameters.

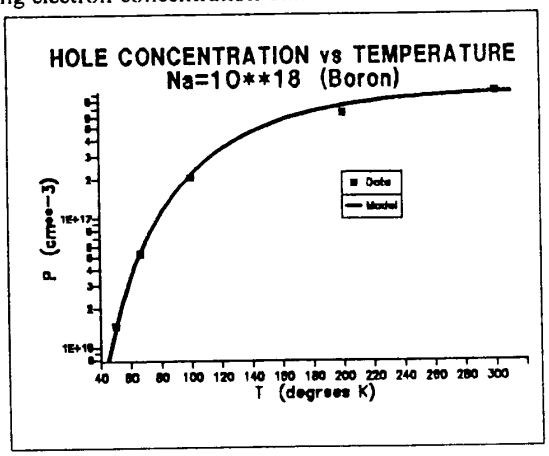

Figure 2. Hole conc. vs. T for Si w/ boron.

Thus, this modeling procedure attempts to fit experimental data with the coefficients ΔE_D, ΔE_A, F_D, and F_A, where the assumption is made that these parameters largely depend only upon concentration, and not very strongly upon temperature. The above least squares method does not work very well when the temperature is too large, since then the argument of the natural logarithm function in Eq. (7) becomes very small ($N_A \simeq p$ at room temperature, and N_D and n are small for a p doped sample); small errors in the data then appear as large errors in the LHS of Eq. (7). Hence, Eq. (7) is mainly useful for fitting data below about $200°K$. However, the lower temperature regime is precisely the region where incomplete ionization is most noticeable, and where the values of ΔE_D, ΔE_A, F_D, and F_A are most crucial. Hence, the difficulty mentioned above is not really very serious.

Other assumptions that should be mentioned here in fitting the above coefficients, is that electron and hole effective mass models from Ref. (11) were used in the standard expressions for N_C and N_V. Secondly, the restriction was made that F_D and F_A should equal 2.0 and 4.0, respectively, at low impurity concentrations. Finally, it should be mentioned that Hall concentration data is uncertain to the extent that the Hall coeffi-

cient factor ρ_H is not accurately known. This quantity equals the ratio between Hall mobility to conductivity mobility, and must be used to deduce carrier concentrations from Hall data. For the Morin and Maita data, we followed their assumptions on the value of ρ_H . For most of Swirhun's very high concentration data, the fraction of ionized impurities varied only slightly with temperature, at least in comparison with samples containing impurity concentrations of 10^{18} cm^{-3} and less. Indeed, much of the variation in Swirhun's data is undoubtedly due to variations in ρ_H, as Swirhun notes in his thesis. Hence, his data was mainly used to pin down the high doping regime where the ionized impurities became largely independent of temperature. F_D and F_A were set equal to zero at this point. Figures 3 and 4 show the plots of the empirical curves found for ΔE_D, ΔE_A, F_D, and F_A, after forcing smooth curves through the results.

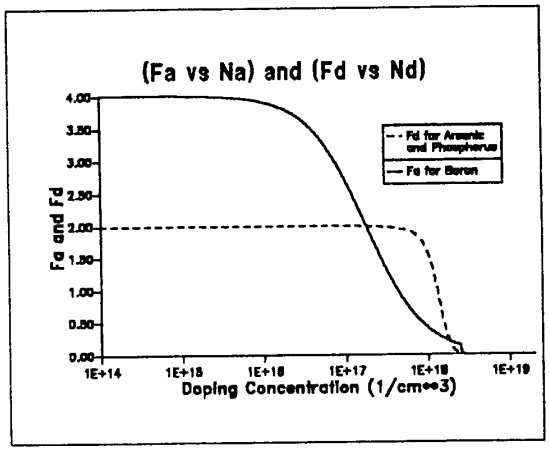

Figure 3. Incomplete ionization model F factors.

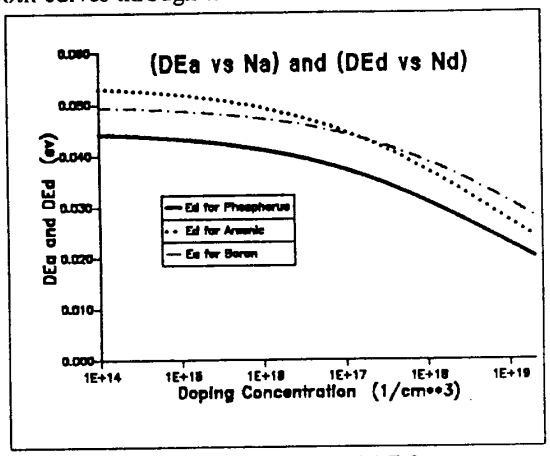

Figure 4. Incomplete ionization model E factors.

For low concentrations, F_A and F_D reduce to the "degeneracy factors" of 4.0 and 2.0, respectively. For high concentrations, these factors equal zero, thereby reflecting the fact that the impurities are fully ionized over the temperature range modeled here of $50 \rightarrow 300°K$. For low concentrations, the ΔE factors reduce fairly closely to the impurity ionization energies found in, for example, Ref. 3 on p. 21.

EXAMPLES OF USE IN SIMULATION

The model described here has been implemented in the device modeling program FIELDAY. Figures 5 and 6 show some results for a simple n-channel MOSFET structure, with Gaussian source and drain profiles that have peak values of 10^{19} cm^{-3} phosphorus. The substrate is assumed to be uniformly doped with 10^{16} cm^{-3} boron. The gate oxide has a 120 Å thickness, and the work function of the gate contact is taken to be that of intrinsic silicon. Simulations were carried out using (a) the incomplete ionization model described here (CJ89), and compared with the physically less accurate cases where (b) all impurities are assumed to be ionized (ALL), and where (c) the low doping model (LOW) of Eqs. (1) and (2) is used indiscriminately for all doping levels, with F_D , F_A , ΔE_D, and ΔE_A taken to be independent of impurity concentration, and given by their values in [3]. As can be seen from Figs. 5 and 6, the differences in these models upon source to drain current shows up at $T = 77°K$, but not at $T = 300°K$, as we should expect, and becomes increasingly important for current magnitude at shorter channel lengths. This latter effect we should also expect, since the source and drain regions, and the transition region from these diffusion pockets to the channel region, is where the three models yield significantly different amounts of ionized impurities. Other simulations that we've carried out reveal that if an LDD structure is used, then the differences between I_{ds} values for the ALL and CJ89 models can easily differ by over 25% at the shorter channel lengths, since there is then a longer transition region in the device with impurity concentrations near 10^{18} cm^{-3}.

Finally, we want to mention the importance of including the time derivatives of N_D^+ and N_A^- in solving current continuity equations for holes and electrons in transient device simulations. More specifically, since $\dot{\rho} = -\nabla \cdot \vec{j}$, and since nearly all conduction occurs in the conduction and valence bands so that $\vec{j} = \vec{j}_n + \vec{j}_p$ (i.e., little conduction occurs in the impurity bands), then we must have that

$$-q\frac{\partial}{\partial t}(n - N_D^+) = -\nabla \cdot \vec{j}_n + R, \qquad (8)$$

$$+q\frac{\partial}{\partial t}(p - N_A^-) = -\nabla \cdot \vec{j}_p - R, \qquad (9)$$

where R has the same value in both equations. Here, we should interpret R as being a recombination-generation (R-G) rate describing the net rate at which the total electron concentration in the conduction and donor impurity bands decrease or increase per unit time. However, even if R is small, we cannot in general ignore $\dot{N}_D$ and $\dot{N}_A$ in Eqs. (8) and (9). Variations in E_{fn} due to a time varying applied potential will change the electron concentration in both the donor and conduction bands. If the time rate of change is not too large, then our model should be suitable for estimate purposes. Based on this model, N_D^+ depends on $(E_{fn} - E_C)/kT$, just as does n in Eq. (3). Hence,

$$\frac{\partial}{\partial t}(n - N_D^+) = \frac{\partial n}{\partial t}\left[1 - \frac{\partial N_D^+}{\partial n}\right]. \tag{10}$$

Using our rather simplistic model, the quantity in brackets above can be shown to be always greater than or equal to one; depending on the value of E_{fn}, this factor can become much larger than 1.0 at low temperatures and for $N_D \simeq 10^{18}$ cm^{-3}. Similar statements hold for the acceptor case.

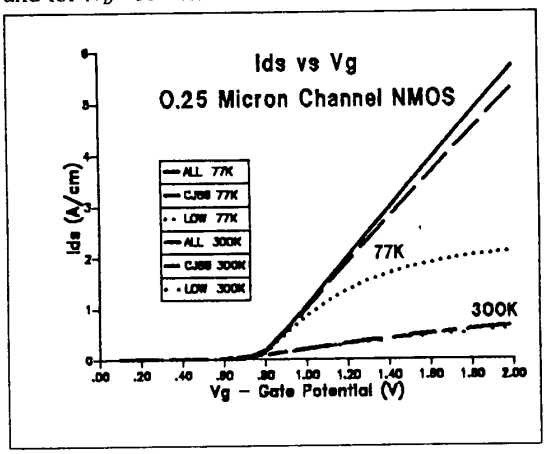

Figure 5. Ids vs Vg for 0.25 micron NMOS device.

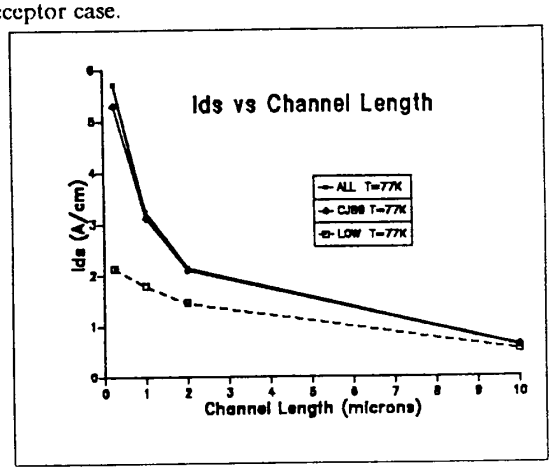

Figure 6. Ids at Vg = 2V, vs. channel length.

ACKNOWLEDGMENT

We thank Dr. F. Stern for his helpful ideas on the model described here.

REFERENCES

[1] D. C. Look, "Statistics of Multicharge Centers in Semiconductors: Applications," *Phys. Rev. B*, vol. 24, pp. 5852-5862, 1981.

[2] R. F. Pierret, *Advanced Semiconductor Fundamentals*, Addison-Wesley, Reading, Mass., 1987, Sec. 4.4.4.

[3] S. M. Sze, *Physics of Semiconductor Devices*, Wiley Interscience, New York, 1981, Sec. 1.4.3.

[4] G. D. Mahan, "Energy gap in Si and Ge: Impurity dependence," *J. Appl. Phys.*, vol. 51, pp. 2634-2646, 1980.

[5] K.-F. Berggren and B. E. Sernelius, "Bandgap narrowing in heavily doped many-valley semiconductors," *Phys. Rev. B*, vol. 24, pp. 1971-1986, 1981.

[6] A. Selloni and S. T. Pantelides, "Electronic Structure and Spectra of Heavily Doped n-Type Silicon," *Phys. Rev. Lett.*, vol. 49, pp. 586-589, 1982.

[7] W. Kuzmicz, "Ionization of Impurities in Silicon," *Solid-State Electron.*, vol. 29, pp. 1223-1227, 1986.

[8] S. E. Swirhun, "Characterization of Majority and Minority Carrier Transport in Heavily Doped Silicon," Phd. thesis, Stanford University, 1987.

[9] F. J. Morin and J. P. Maita, "Electrical Properties of Silicon Containing Arsenic and Boron," *Physical Review*, vol. 96, pp. 28-35, 1954.

[10] A. H. Marshak and C. M. Van Vliet, "Electrical Current and Carrier Density in Degenerate Materials with Nonuniform Band Structure," *Proc. IEEE*, vol. 72, pp. 148-164, 1984.

[11] F. H. Gaensslen and R. C. Jaeger, "Temperature Dependent Threshold Behavior of Depletion Mode MOSFET's," *Solid-St. Electron.*, vol. 22, p. 423, 1979.