Theoretical Limit of Low Temperature Subthreshold Swing in Field-Effect Transistors

Arnout Beckers®, Farzan Jazaeri®, and Christian Enz®, Fellow, IEEE

Abstract—This letter reports a temperature-dependent limit for the subthreshold swing in MOSFETs that deviates from the Boltzmann limit at deep-cryogenic temperatures. Below a critical temperature, the derived limit saturates to a value that is independent of temperature and proportional to the characteristic decay of a band tail. The proposed expression tends to the Boltzmann limit when the decay of the band tail tends to zero. Since the saturation is universally observed in different types of MOSFETs (regardless of dimension or semiconductor material), this suggests that an intrinsic mechanism is responsible for the band tail.

Index Terms—Band tail, cryogenic, MOSFET, modeling, subthreshold slope, subthreshold swing.

I. INTRODUCTION

The Boltzmann limit of the subthreshold swing in FETs, SS = (kBT/q)ln 10, predicts at room temperature the well-known ≈ 60 mV/dec, and at deep-cryogenic temperatures (<≈ 50 K) an almost ideal, step-like switch (kBT/q is the thermal voltage). However, the measurements in FETs at deep-cryogenic temperatures reach merely ≈ 11 instead of 0.8 mV/dec at 4.2 K [1]–[6], ≈ 9 mV/dec instead of 20 µV/dec at 100 mK [7], and ≈ 7 mV/dec instead of 4 µV/dec at 20 mK [8]. As shown in Fig. 1, this degradation is measured in structurally different FETs, operating in subthreshold at both low and high drain-source voltage (VDS) and for various technologies: mature and advanced bulk and FDSOI (Fully-Depleted Silicon-On-Insulator) MOSFETs [4]–[14], FinFETs [15], [16], gate-all-around Si nanowire FETs [17], junctionless FETs [18], [19], SiGe FETs [20], InP HEMTs [21], SiC FETs [22], etc. Figure 1 highlights this measured trend, deviating from the Boltzmann limit below a critical temperature, and then saturating to a value depending on the technology. The difference between the measured SS and the Boltzmann limit is referred to as excess SS. The additional power that the FET consumes at deep-cryogenic temperatures due to the excess SS is a crucial metric for the realization of quantum processors in silicon [23]–[30] and for assessing the benefits of temperature scaling as an alternative to traditional scaling [9], [10], [31].

It is simply not possible to explain the saturation of SS(T) using the Boltzmann limit. Indeed, the Boltzmann limit is linear in T. Its slope versus T in the plot in Fig. 1 is proportional to the slope factor m0 [SS = m0(kBT/q)ln 10], where m0 = 1 + Cdepl/Cox, with Cox the gate-oxide capacitance, and Cdepl the depletion capacitance. The slope factor m0 is limited to 2 since Cdepl < Cox which allows to explain the measurements only down to ≈ 50 K. Including the interface-trap capacitance CIT ∝ qNIT (NIT is the number of interface states per unit area) in the slope factor (m = m0 + qNIT/Cox), does not help to model the behavior below ≈ 50 K, since this only further increases the linear slope of SS versus T.

Furthermore, this approach has led to unrealistically high NIT at deep-cryogenic temperatures. Typical NIT values that have been reported in the literature are in the order of 1013 − 1014 cm−2 at 4.2 K [16], [18], [19], and 1015 cm−2 at 20 mK [8]. The values at 4.2 K are still possible in principle. The values at 20 mK, however, exceed 7 × 1014 cm−2 corresponding to the number of atomic lattice sites per unit area in silicon. Furthermore, it should be emphasized that the Boltzmann limit leads to a singularity in NIT near 0 K.

Recently, relying on numerical simulations Bohuslavský et al. demonstrated that an exponential band tail and Fermi-Dirac statistics leads to saturation of SS at deep-cryogenic temperatures [32], [33]. The presence of a band tail in FDSOI FETs was explained by a combination of crystalline disorder, strain,
residual impurities, etc. However, the saturation of SS(T) has been measured in older technologies as well, before strain and nanometer dimensions were introduced that lead to disorder. Since SS(T) is fairly independent of technology, this suggests that the extent of the band tail is fairly independent of technology, which points in the direction of an intrinsic mechanism being responsible for blurring the band edges (e.g., electron-phonon scattering, electron-electron and electron-hole interactions, finite crystalline periodicity, etc.) and to a lesser extent extrinsic mechanisms (impurities, disorder, defects, etc.) [34].

II. Revisited Theoretical Limit

The total drain current in subthreshold can be approximated by $I_{DS} = q(W/L)\mu (k_B T/q)(n_D - n_S)$, adopting the unified charge-controlled model that is valid for bulk, SOI, FinFET, and other multigate FETs including quantum effects [35-37] ($q$ is the electron charge, $W/L$ the width-over-length ratio of the transistor gate, $\mu$ the free-carrier mobility assumed constant along the channel, and $n_D$ and $n_S$ are the electron densities at the drain and source sides, assuming an n-channel FET without loss of generality). Hence, $SS = \partial V_GS/\partial \log I_{DS}$ can be expressed as $m[(n_D - n_S)/(\partial \psi_S/\partial \psi_S - \partial \psi_S/\partial \psi_S)]\ln 10$, where $V_GS$ is the gate-to-source voltage, $V = \partial V_GS/\partial \psi_S = 1 + (C_{depl} + C_{it})/C_{ox}$, and $\psi_S$ is the electrostatic potential at the surface compared to the bulk [Fig. 2(a)]. The electron density in a conduction-band tail [Fig. 2(b)] is described by:

$$n = \int_{-\infty}^{E_{c,s}} \text{DOS}(E)\exp\left(\frac{E - E_{c,s}}{W_t}\right) f(E)dE,$$  

where $E_{c,s}$ is the conduction-band energy of the sharp band edge at the surface, $W_t$ is the characteristic decay of an exponential band tail in the bandgap, and $f(E)$ is the Fermi-Dirac function. For simplicity, since SS will not depend on the exact value of $\text{DOS}(E)$, we assume that $\text{DOS}(E_{c,s})$ can be given by the conduction-band DOS in 2-D: $N_{c}^{2D} = g_s m^*/(\pi \hbar^2)$, where $g_s = 2$ is the degeneracy factor, $m^* = 0.19 m_e$ is the effective mass in silicon (assumed temperature independent), $m_e$ the electron mass, and $\hbar$ the reduced Planck constant. The solution of integral (1) takes the form of a Gauss hypergeometric function ($F_1(a, b; c; z)$) [38]:

$$n = N_c^{2D} W_t F_1(1, \theta; \theta + 1; z),$$  

where $\theta = k_B T/W_t$, $z = -\exp[\left(\frac{E_{c,s} - E_{F,n}}{k_B T}\right)]$ and $E_{F,n} = E_F - q V$ is the quasi-Fermi energy of electrons and $V$ is the channel voltage. The band diagram in Fig. 2(a) shows that $E_F = E_F^0 - E_g - q V$, where $E_F^0$ is the conduction-band energy in thermal equilibrium, $E_g$ the bandgap, and $F_\Phi = (k_B T/q) \ln (N_A/n_l)$ the Fermi potential with $N_A$ the doping concentration in the MOSFET body and $n_l$ the intrinsic carrier concentration.

The expression given for $\Phi_\Phi$ assumes Boltzmann statistics, which has been verified at deep-cryogenic temperatures in case there is no band tail [29]. If a band tail is present, Fermi-Dirac statistics ought to be used like in (1) because the Fermi level can lie in the band tail which violates $E - E_{F,n} \gg 3k_B T$. In principle, the valence-band tail should be taken into account in the derivation of $\Phi_\Phi$. However, in the charge-neutrality condition in the p-type bulk ($p = N_A$), we can assume that there is no valence-band tail, hence Boltzmann statistics can be used to arrive at $\Phi_\Phi = (k_B T/q) \ln (N_A/n_l)$. Including the valence-band tail in $p$ would lead to a different value of $\Phi_\Phi$, which shifts the electron current, but does not change its slope. The same argument holds when dopant freezeout is included ($p = N_A^2$) [39]. For p-channel FETs, the roles of the conduction-band and valence-band tails would be reversed; the valence-band tail being responsible for the saturation of SS, and the conduction-band tail negligible in the computation of SS.

The bandgap is only slightly temperature dependent in the cryogenic regime [40]. Similar to a valence-band tail and dopant freezeout, bandgap widening will shift the electron current, but not change its slope.

Using $\psi_S = -\left(E_{c,s} - E_F^0\right)/q$, it follows that $E_{F,n} - E_{c,s} = q \psi_S - E_g/2 - q \Phi_\Phi - q V$. The latter can be inserted in (2) to yield $n$ as a function of $\psi_S$, where $z = -\exp\left[-q \psi_S/(k_B T)\right]$. $\psi_S' = \psi_S - \psi_S^*$, and $\psi_S^* = E_g/(2q) + \Phi_\Phi$. The defined $\psi_S^*$ depends only on $T$ and $N_A$ at a fixed $V_{DS}$. Note that for $\psi_S$ in subthreshold, ranging from 0 (flatband) to 2$q\Phi_\Phi + V$ (threshold), $\psi_S'$ is always negative. The first derivative of a hypergeometric function $F_1(a, b; c; z)$ is given by $(ab/c)F_1(a + 1, b + 1; c + 1; z)$ [41]. Differentiating (2) with respect to $\psi_S$ (applying the chain rule for $z$), we find that

$$\frac{\partial n}{\partial \psi_S} = -q z N_c^{2D} F_1(2, \theta + 1; \theta + 2; z).$$

Inserting (2) and (3) in the expression for SS, gives:

$$SS = m\left(\frac{k_B T}{q}\right) \ln 10 \times A[z(\psi_S, V), T, W_t],$$

where $A$ is given by

$$\left[\frac{\theta^{-1} F_1(1, \theta; 1; \theta + 1; z)}{\theta + 1} - 1\right]^{-1} F_1(2, \theta + 1; 1; \theta + 2; z) = \frac{V_{DS}}{V_{VDS}}.$$  

Expression (4)-(5) is plotted in Fig. 3 versus $T$ and $\psi_S$ for different $W_t$ together with the Boltzmann limit. As shown in Fig. 3, (i) SS rolls off from the Boltzmann limit and saturates at deep-cryogenic temperatures, (ii) the saturation value of SS increases with $W_t$, (iii) the critical temperature at which SS starts to deviate from the Boltzmann limit increases with $W_t$.

Expression (4)-(5) tends to the Boltzmann limit (i) at high temperatures ($k_B T \gg W_t$ or $\theta \to \infty$), and (ii) when the bandtail decay tends to zero ($W_t \to 0$ or $\theta \to \infty$). In both cases, $\theta \approx \theta + 1 \approx \theta + 2$ in (5). Using the relation $F_1(a, b; b; z) = (1 - z)^{-a}$ (which is valid for all $b$ [41]) on both $F_1$ in (5), it follows that $A$ tends to $(z - 1)/z$. Since $\psi_S^*$ in $z$ is always negative in subthreshold, $A$ tends to 1. The Boltzmann limit is then obtained in (4) for cases (i) and (ii).

Figures 4(a) and 4(b) show the measured transfer characteristics in a large bulk silicon, n-channel MOSFET
cryogenic temperatures 

\[ \text{SS} \approx \frac{W_i}{q} \ln 10 \] 

where \( W_i \) is in Joules. Applying one of Euler’s linear transformations for hypergeometric functions,

The above result confirms the saturation value of \( SS \) that was apparent in the numerical simulations by Bohuslavskyi et al. [32], [33]. The revised limit in (4)-(5) and saturation value in (8) are valid for bulk, SOI, FinFET, nanowire, and other multigate FETs, provided that \( m \) accounts for enhanced electrostatic control. The horizontal planes in Fig. 3 indicate the saturation values for increasing \( W_t \). The critical temperature for which \( SS \) deviates from the Boltzmann limit can be estimated as \( T_{crit} = W_i/k_B \).

In Fig. 5, \( T_{crit} \) is about 46 K. Similarly, \( T_{crit} \) provides a simple method to obtain \( W_i \) from dc measurements by plotting \( SS \) versus \( T \). Furthermore, (8) demonstrates that the deep-cryogenic subthreshold performance of FETs is determined by \( W_i \) and \( N_{it} \) (in \( m \)). A more reasonable \( N_{it} \) can be extracted at sub-Kelvin temperatures by using (8) instead of the Boltzmann limit. The singularity of \( N_{it} \) near 0 K is also avoided. Since (8) is independent of \( V_{DS} \), the slightly higher \( SS \) in Fig. 1 at \( V_{DS} = 0.9 \text{ V} \) compared to \( V_{DS} = 5 \text{ mV} \) is a high-field effect.

IV. CONCLUSION

An analytical expression for the saturating \( SS(T) \) is derived from room down to sub-Kelvin temperature. When the thermal energy becomes smaller than the band-tail extension \( (W_i) \), the revised \( SS(T) \) limit follows the temperature-independent \( m(W_i/q) \ln 10 \) rather than \( m(k_BT/q) \ln 10 \). The revised limit demonstrates that a perfect MOS switch \( (SS = 0) \) cannot be obtained in the presence of a band tail. The problem of extracting anomalously high interface-trap density at deep-cryogenic temperatures is solved by using \( m(W_i/q) \ln 10 \).

REFERENCES


