# The effect of density-of-state tails on band-to-band tunneling: Theory and application to tunnel field effect transistors

S. Sant, and A. Schenk

Citation: Journal of Applied Physics **122**, 135702 (2017); doi: 10.1063/1.4994112
View online: http://dx.doi.org/10.1063/1.4994112
View Table of Contents: http://aip.scitation.org/toc/jap/122/13
Published by the American Institute of Physics

# The effect of density-of-state tails on band-to-band tunneling: Theory and application to tunnel field effect transistors

S. Sant[a] and A. Schenk[b]
*Integrated Systems Laboratory, ETH Zurich, Gloriastr. 35, 8092 Zurich, Switzerland*

(Received 4 July 2017; accepted 1 September 2017; published online 4 October 2017)

It is demonstrated how band tail states in the semiconductor influence the performance of a Tunnel Field Effect Transistor (TFET). As a consequence of the smoothened density of states (DOS) around the band edges, the energetic overlap of conduction and valence band states occurs gradually at the onset of band-to-band tunneling (BTBT), thus degrading the sub-threshold swing (SS) of the TFET. The effect of the band tail states on the current-voltage characteristics is modelled quantum-mechanically based on the idea of zero-phonon trap-assisted tunneling between band and tail states. The latter are assumed to arise from a 3-dimensional pseudo-delta potential proposed by Vinogradov [1]. This model potential allows the derivation of analytical expressions for the generation rate covering the whole range from very strong to very weak localization of the tail states. Comparison with direct BTBT in the one-band effective mass approximation reveals the essential features of tail-to-band tunneling. Furthermore, an analytical solution for the problem of tunneling from continuum states of the disturbed DOS to states in the opposite band is found, and the differences to direct BTBT are worked out. Based on the analytical expressions, a semi-classical model is implemented in a commercial device simulator which involves numerical integration along the tunnel paths. The impact of the tail states on the device performance is analyzed for a nanowire Gate-All-Around TFET. The simulations show that tail states notably impact the transfer characteristics of a TFET. It is found that exponentially decaying band tails result in a stronger degradation of the SS than tail states with a Gaussian decay of their density. The developed model allows more realistic simulations of TFETs including their non-idealities. *Published by AIP Publishing.* [http://dx.doi.org/10.1063/1.4994112]

## I. INTRODUCTION

Tunnel Field Effect Transistors (TFETs) are considered as a low-power alternative to the Metal-Oxide-Semiconductor Field Effect Transistor (MOSFET) because of their theoretical capability to exhibit sub-thermal sub-threshold swing (SS).[2,3] Band tail states have captured renewed interest among TFET designers because of their detrimental influence on the TFET characteristics.[4,5] The operation mode of a TFET is electron-hole pair generation by band-to-band tunneling (BTBT) instead of thermionic injection in a MOSFET. BTBT starts when the conduction band (CB) edge energetically aligns with the valence band (VB) edge. When both band edges are abrupt (as in an ideal semiconductor), the energetic overlap takes place abruptly which results in steeper switching of the device [Fig. 1(a)]. Band tails close to either of the bands represent a gradual decay of the electron/hole density of states (DOS) into the gap. Therefore, the energetic alignment of the two bands in the presence of band tails takes place gradually which degrades the SS of the TFET [Fig. 1(b)]. Besides DOS tails, the SS of TFETs is mainly affected by Shockley-Read-Hall (SRH) generation leakage and trap-assisted tunneling at interfaces[5] and in bulk regions.[6] The latter can be viewed as a field-enhanced multi-phonon process which depends on the concentration of active defects and temperature. Another field-enhanced generation process is tunnel-assisted impact ionization[7] which was recently identified as intrinsic limitation to the SS of TFETs.[8]

Band tails can originate from random dopant placement or the presence of defects, which gives rise to electronic states in the forbidden gap close to the band edges.[9–11,14] Random placement of the dopant atoms creates, on ionization, randomly placed charge centers in the crystal lattice. This forms the band tail states which are highly localized in the warped region, but are strongly coupled to the nearest band edge. The origin of band tails has been studied in detail by many authors using different approaches. Kane derived

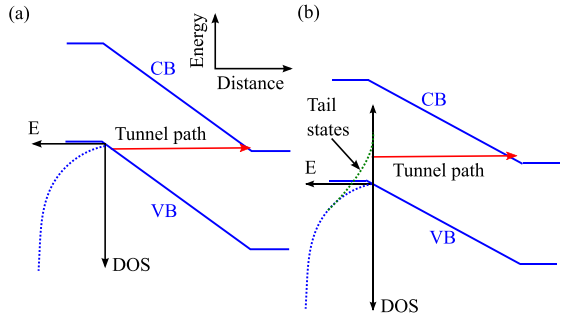


FIG. 1. Schematic band edge diagram of a semiconductor at the onset of BTBT (a) in the absence of band tails and (b) when valence band tails are present. In the latter case, the onset of BTBT is less sharp because of the gradual increase of the DOS.

[a]Electronic mail: sasant@iis.ee.ethz.ch
[b]Electronic mail: schenk@iis.ee.ethz.ch

the density of tail states originating from random dopant fluctuation using the Thomas-Fermi approximation.[15] Halperin and Lax[20] used the minimum counting method and computed the DOS tails numerically. Efros[21] introduced an optimal fluctuation method to derive the density of tail states deep in the tail.

In the experimental analysis of the effect of tail states, the DOS in the band gap is approximated by an exponential ($A\exp(-(E_c - E_t)/\eta)$) or by a Gaussian ($A\exp(-(E_c - E_t)^2/\eta^2)$) function.[22] The characteristic energy $\eta$ is obtained from photoluminescence (PL) measurements of the semiconductor. The value of $\eta$ for bulk InAs was found to be 7 meV (Refs. 23 and 24) for unintentionally doped InAs while being 25 meV (Ref. 25) in InSb when fitted to an exponential. For n-doped GaAs, $\eta$ ranges from 20 meV–25 meV.[26] Conductivity measurements, when fitted to a Gaussian decay mode, yielded the value of $\eta$ for GaAs to be around 60 meV.[27]

The effect of band tails on the transfer characteristics of a nanowire TFET has been analyzed by Khayer and Lake[4] using a Non-equilibrium Green's Function approach. Their study revealed a strong degradation of the TFET characteristics in the presence of band tails. A characteristic tail energy of $\eta = 25$ meV was found to increase the SS by a factor of four.

In Sec. II, we present a quantum-mechanical treatment to analytically calculate the tunnel rate of electrons from VB tail states into the CB. The entire range from very strong to very weak localization is considered as well as the modifications for tunneling from continuum states of the non-ideal VB DOS into CB states. For all cases, the essential differences compared to ideal direct BTBT are worked out. The analytical results are then used to modify a semi-classical BTBT rate which is implemented in a Finite Element Method (FEM) based TCAD simulator. The impact of various parameters of the model on the TFET performance is studied for the case of a nanowire transistor in Sec. III. A summary and the conclusions of the study are given in Secs. IV and V, respectively. Appendixes A–D contain the details of the analytical derivations.

## II. QUANTUM-MECHANICAL MODEL FOR TUNNELING BETWEEN TAIL AND BAND STATES

In the following, after introducing the DOS tail model, three analytical models for the rate of tail-to-band tunneling in a constant electric field will be derived based on the degree of localization of the tail states (see Table I). They are denoted *model-0* in the case of strongly localized tail states neglecting their field broadening, *model-1* in the case of field-broadened strongly localized tail states, and *model-2* in the case of field-broadened weakly localized tail states.

TABLE I. Summary of analytical models of the tail-to-band tunnel rate.

| Model | Localization | Field broadening |
|---|---|---|
| Model-0 | Strong | Neglected |
| Model-1 | Strong | Included |
| Model-2 | Weak | Included |

### A. DOS tail model

Random dopant distributions and crystal defects in the semiconductor give rise to localized states with different energies inside the band gap which in the aggregate form a band tail. According to Kane's theory of band tail states,[15] the DOS of a semiconductor in the presence of random dopant fluctuations takes the form

$$\varrho_{t,e/h}(E) = \frac{(2m_{t,e/h})^{3/2}}{2\pi^2\hbar^3}\sqrt{\eta}\,Y_{g/e}(E/\eta)\,, \tag{1}$$

where $m_{t,e/h}$ is the effective mass of a carrier in the tail state, $\hbar$ is reduced Planck's constant, and $\eta$ is the characteristic energy of the band tail. In Eq. (1), the energy E counts from the respective band edge. In the case of Gaussian tails, the function $Y(E/\eta)$ is given by

$$Y_g(x) = \frac{1}{\sqrt{\pi}}\int_{-\infty}^{x} d\zeta\,\sqrt{x-\zeta}\,e^{-\zeta^2} \tag{2}$$

which in the limit $x \to \infty$ turns into

$$Y_g(x) \to \frac{1}{2^{5/2}x^{3/2}}\,e^{-x^2}\,. \tag{3}$$

Therefore, if $E \gg \eta$, $Y_g(E/\eta)$ can be approximated by Eq. (3) and the tail DOS becomes

$$\varrho_{t,e/h}(E) = \frac{(m_{t,e/h})^{3/2}}{4\pi^2\hbar^3}\frac{\eta^2}{E^{3/2}}\exp\left[-(E/\eta)^2\right]. \tag{4}$$

Note that the approximation (4) cannot be used if $|E| < \eta$ or if $E$ is an energy in the band. In this case, Eq. (1) has to be applied which requires a numerical integration, or Eq. (3) is empirically modified such that it approximates $Y_g(x)$ up to $x = 0$. A simple, but efficient modification is given by

$$Y_g(x) \to \frac{e^{-x^2}}{2^{5/2}(x^{3/2}+s)} \tag{5}$$

with $s = 0.566$. The modified expression for the tail DOS then reads

$$\varrho_{t,e/h}(E) = \frac{(m_{t,e/h})^{3/2}}{4\pi^2\hbar^3}\frac{\sqrt{\eta}}{[(E/\eta)^{3/2}+s]}\exp\left[-(E/\eta)^2\right]. \tag{6}$$

In the case of exponential tails, the Gaussian function in the integrand of Eq. (2) is replaced by an exponential function and $Y(E/\eta)$ becomes

$$Y_e(x) = \frac{1}{2}\int_{-\infty}^{x} d\zeta\,\sqrt{x-\zeta}\,e^{-|\zeta|} \tag{7}$$

which yields the correct ideal DOS in the limit $\eta \to 0$. For $x < 0$, the integral can be calculated exactly:

$$Y_e(x) = \frac{1}{4\sqrt{\pi}}e^{-|x|}, \qquad x < 0. \tag{8}$$

For $x > 0$ (energies in the band), Eq. (7) has to be solved numerically. Obviously, $Y_e(x) \to \sqrt{x}$ for $x \to \infty$.

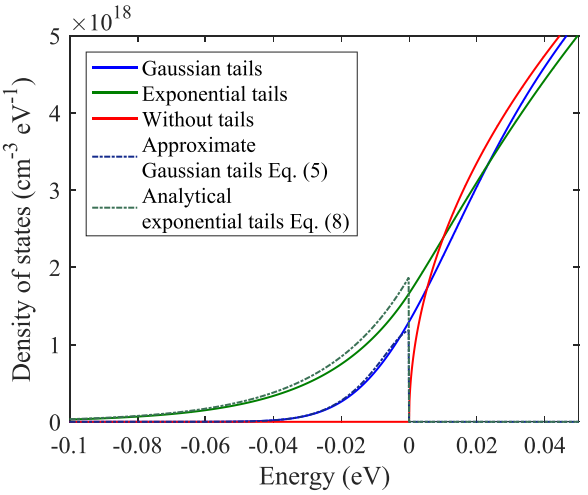


FIG. 2. Comparison of the conduction band DOS in InAs with Gaussian and exponential tails calculated by the exact expression (1) and by various approximations.

All functions defining Gaussian and exponential tails are compared in Fig. 2.

### B. Tunneling from/into tail states ignoring their life time Broadening

In this subsection, tunneling from tail states at the VB into the ideal CB is considered. For the rate derivation, it is assumed that they are sufficiently localized such that the pseudo-$\delta$-potential model[1,16] can be used which enables to express the transition rate between localized states near the VB (energy $E_t$, spatial density $1/(2\pi r_0^3)$) and a CB state by[17]

$$D_{t,c}(E,E',E_t) = 8\pi \frac{\mathcal{P}}{\tilde{E}^2} E_t^2 r_0^3 \varrho_t(\tilde{E},E_t)\,\varrho_c(E')\,. \tag{9}$$

Here, $\mathcal{P}$ is the Cauchy principal value of integrals over $\tilde{E} = E + E_g$, and $\varrho_t(\tilde{E},E_t)$ denotes the density of the localized single-level states

$$\varrho_t(\tilde{E},E_t) = \frac{1}{2\pi r_0^3}\delta(\tilde{E} - E_t)\,. \tag{10}$$

The energy level $E_t$ of the tail state is measured from the VB edge. The different energy variables in Eq. (9) and their meaning in the derivation below are presented in Fig. 3.

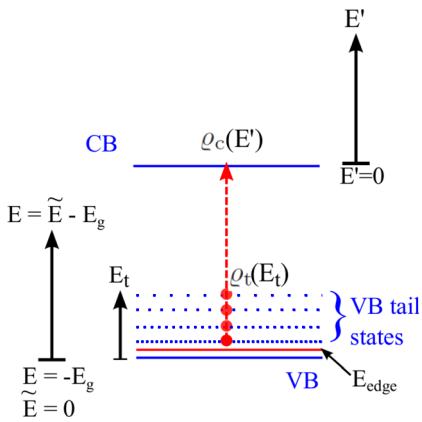


FIG. 3. Representation of the energy variables used in Eq. (9) and thereafter.

For Eq. (10), the field effect on the localized state was neglected.[17] It will be included in Subsection II C. Note that Eq. (9) is a special case of Eq. (27) which is derived in Appendix A. The "effective" mass $m_t$ of the localized electron is simply related to its localization radius $r_0$ by

$$E_t = \frac{\hbar^2}{2m_t r_0^2}\,, \tag{11}$$

which can be viewed as fitting of the localization radius $r_0$. The mass $m_t$ appears as a result of the single-band envelope method applied to the localized state, a method limited to a single, parabolic, and isotropic band. This parameter takes account of the presence of heavy and light holes and the anisotropy of their bands. Deeper tail states might even be affected by the conduction bands. Hence, $m_t$ becomes a parameter of the analytical model. Table II lists the localization radius $r_0$ for various values of $m_t$ at three values of $E_t$.

The tail DOS Eq. (1) (with $E$ replaced by $-\tilde{E}$) is assumed to be a dense ladder of single-level DOSs Eq. (10) and will be composed with weight functions $w(E_t)$ as

$$\varrho_{t,h}(\tilde{E}) = \int_{E_{edge}}^{E_g} dE_t w(E_t)\varrho_t(\tilde{E},E_t)\,. \tag{12}$$

The integration over tail states is restricted to $E_{edge} < \tilde{E} < E_g$ (see Fig. 3). The lower limit $E_{edge}$ separates localized states from continuum states, and it is assumed that $0 < E_{edge} < \eta$. Thus, $E_{edge}$ plays the same role as the "mobility edge"[18] in transport. The mass $m_{t,h}$ from Kane's DOS model could be set to the hole mass $m_h$, as done by Kane,[14] or to $m_t$ which is preferred here (see discussion above). The model has then just *one* fitting parameter to account for the unknown electronic structure of the tail states. However, its value cannot differ vastly from that of a hole mass.

The weight function $w(E_t)$ immediately follows from inserting the DOS expressions into the above equation:

$$w(E_t) = \frac{\sqrt{\eta}}{\pi E_t^{3/2}} Y(E_t/\eta)\,. \tag{13}$$

In the case of Gaussian tails, it can be simplified using (5) to

$$w(E_t) = \frac{\sqrt{\eta}\exp\left(-E_t^2/\eta^2\right)}{2^{5/2}\pi E_t^{3/2}\left[(E_t/\eta)^{3/2} + s\right]}\,. \tag{14}$$

TABLE II. Localization radius for various values of the effective mass $m_t$.

| $E_t = 0.01$ eV | | $E_t = 0.025$ eV | | $E_t = 0.05$ eV | |
|---|---|---|---|---|---|
| $m_t(m_0)$ | $r_0(nm)$ | $m_t(m_0)$ | $r_0(nm)$ | $m_t(m_0)$ | $r_0(nm)$ |
| 0.001 | 87.34 | 0.001 | 55.24 | 0.001 | 39.06 |
| 0.01 | 27.62 | 0.01 | 17.47 | 0.01 | 12.35 |
| 0.025 | 17.47 | 0.025 | 11.05 | 0.025 | 7.81 |
| 0.05 | 12.35 | 0.05 | 7.81 | 0.05 | 5.52 |
| 0.1 | 8.73 | 0.1 | 5.52 | 0.1 | 3.91 |
| 1 | 2.77 | 1 | 1.75 | 1 | 1.23 |

The transition rate between a tail state with energy $E$ and a CB state with energy $E'$ becomes

$$D_{\mathrm{t,c}}(E,E') = \int_{E_{\mathrm{edge}}}^{E_{\mathrm{g}}} dE_{\mathrm{t}} w(E_{\mathrm{t}}) D_{\mathrm{t,c}}(E,E',E_{\mathrm{t}}) = 8\pi \left(\frac{\hbar^2}{2m_{\mathrm{t}}\tilde{E}}\right)^{\frac{3}{2}} \varrho_{\mathrm{t,h}}(\tilde{E})\varrho_{\mathrm{c}}(E') \tag{15}$$

with

$$\varrho_{\mathrm{t,h}}(\tilde{E}) = \left(\frac{2m_{\mathrm{t}}}{\hbar^2}\right)^{\frac{3}{2}} \frac{\sqrt{\eta}}{2\pi^2} Y(\tilde{E}/\eta)\,. \tag{16}$$

In a constant electric field $F$ and assuming a parabolic dispersion for the CB, the CB DOS $\varrho_{\mathrm{c}}(E')$ has the form[19]

$$\varrho_{\mathrm{c}}(E') = \frac{\sqrt{8m_c^3}}{4\pi\hbar^3}\sqrt{\hbar\theta_{\mathrm{c}}}\,\mathcal{F}\left(-\frac{E'}{\hbar\theta_{\mathrm{c}}}\right) \tag{17}$$

$$\text{with} \quad \mathcal{F}(x) = Ai'(x)^2 - xAi(x)^2$$

$$\text{and} \quad \hbar\theta_{\mathrm{c}} = \left(\frac{e^2\hbar^2F^2}{2m_{\mathrm{c}}}\right)^{1/3}.$$

Here, $e$ is the electron charge and $Ai(x)$ denotes the Airy function. The total emission rate from all tail states with energy $E$ into the CB is given by

$$G_{\mathrm{tc}} = \frac{(eF)^2 z_{\mathrm{cv}}^2}{\hbar} \int_{E_{\mathrm{edge}}-E_{\mathrm{g}}}^{0} dE \int_{-\infty}^{\infty} dE'\, D_{\mathrm{t,c}}(E,E')\delta(E-E')\,, \tag{18}$$

where $z_{\mathrm{cv}}$ is the interband transition matrix element[29–32] $z_{\mathrm{cv}}^2 = \hbar^2/(4m_{\mathrm{r}}E_{\mathrm{g}})$ with the reduced effective mass $m_{\mathrm{r}} = m_{\mathrm{c}}m_{\mathrm{v}}/(m_{\mathrm{c}}+m_{\mathrm{v}}) = m_{\mathrm{c}}m_{\mathrm{v}}/m_{\Sigma}$. For the completion of the band-to-band process, it is assumed that the thermionic emission step from the VB to the tail state is very fast and, therefore, not rate limiting. Then, after inserting (15) with (4) and (17) into Eq. (18), the generation rate via Gaussian tail states becomes

$$G_{\mathrm{tc}} = \frac{(eF)^2\sqrt{\eta}\sqrt{\hbar\theta_c}}{\sqrt{2}\pi^2\hbar^2 E_{\mathrm{g}}} \frac{m_{\mathrm{c}}^{3/2}}{m_{\mathrm{r}}} \times \int_{E_{edge}}^{E_g} \frac{d\tilde{E}}{(\tilde{E})^{3/2}} Y\left(-\frac{\tilde{E}}{\eta}\right)\mathcal{F}\left(\frac{E_{\mathrm{g}}-\tilde{E}}{\hbar\theta_c}\right). \tag{19}$$

Since $E_{\mathrm{g}} \gg \hbar\theta_{\mathrm{c}}$, the function $\mathcal{F}$ can be replaced by its asymptotic limit for large positive arguments $\mathcal{F}(x) \to \exp(-4x^{3/2}/3)/(8\pi x)$. If $Y$ is replaced by approximation (5), Eq. (19) simplifies to

$$G_{\mathrm{tc}} = \frac{(eF)^3\sqrt{\eta}}{64\sqrt{2}\pi^3\hbar E_{\mathrm{g}}^2}\frac{m_{\mathrm{c}}}{m_{\mathrm{r}}} \times \int_{E_{\mathrm{edge}}}^{E_{\mathrm{g}}} d\tilde{E}\, \frac{\exp\left[-\frac{\tilde{E}^2}{\eta^2} - \frac{4}{3}\left(\frac{E_{\mathrm{g}}-\tilde{E}}{\hbar\theta_{\mathrm{c}}}\right)^{\frac{3}{2}}\right]}{\tilde{E}^{3/2}\left[(\tilde{E}/\eta)^{3/2}+s\right]}. \tag{20}$$

In order to demonstrate what distinguishes the tunnel rate (20) from the rate of band-to-band tunneling (BTBT), the remaining integral is calculated analytically. The integrand is dominated by the overlap of the steep DOS tail of the VB and the so-called Franz-Keldysh tail of the CB. Since the product of both results in a sharply bell-shaped curve, one can determine its peak position $\tilde{E} = \Delta$ approximately and map the integrand to a Gaussian bell curve. This leads to

$$\int_{E_{\mathrm{edge}}}^{E_{\mathrm{g}}} d\tilde{E} \frac{e^{-\frac{\tilde{E}^2}{\eta^2} - \frac{4}{3}\left(\frac{E_{\mathrm{g}}-\tilde{E}}{\hbar\theta_{\mathrm{c}}}\right)^{\frac{3}{2}}}}{\tilde{E}^{3/2}\left[(\tilde{E}/\eta)^{3/2}+s\right]} \approx \frac{e^{-\frac{\Delta^2}{\eta^2} - \frac{4}{3}\left(\frac{E_{\mathrm{g}}-\Delta}{\hbar\theta_{\mathrm{c}}}\right)^{\frac{3}{2}}}}{\Delta^{3/2}\left[(\Delta/\eta)^{3/2}+s\right]} \int_{-\infty}^{\infty} d\epsilon\, e^{-\frac{\epsilon^2}{\gamma^2}} \tag{21}$$

with

$$\Delta \approx \frac{\sqrt{E_{\mathrm{g}}}\eta^2}{(\hbar\theta_{\mathrm{c}})^{3/2}} \quad \text{and} \quad \gamma \approx \eta\,. \tag{22}$$

These expressions rely on the assumption $\eta \ll E_{\mathrm{g}}$. The generation rate finally takes the form

$$G_{\mathrm{tc}} = \frac{(eF)^3\,\eta^{3/2}}{64\sqrt{2}\pi^3\hbar E_{\mathrm{g}}^2}\frac{m_{\mathrm{c}}}{m_{\mathrm{r}}} \times \frac{\exp\left[-\frac{\Delta^2}{\eta^2} - \frac{4}{3}\left(\frac{E_{\mathrm{g}}-\Delta}{\hbar\theta_{\mathrm{c}}}\right)^{\frac{3}{2}}\right]}{\Delta^{3/2}\left[(\Delta/\eta)^{3/2}+s\right]}. \tag{23}$$

When the one-band effective-mass approximation and the WKB limit are applied to compute the generation rate of direct BTBT (with ideal DOS), one obtains[19]

$$G_{\mathrm{BTB}} = \frac{(eF)^3}{64\pi\hbar E_{\mathrm{g}}^2}\exp\left[-\frac{4}{3}\left(\frac{E_{\mathrm{g}}}{\hbar\theta_{\mathrm{r}}}\right)^{\frac{3}{2}}\right]. \tag{24}$$

Comparing the last two equations, three differences become obvious: (i) the tunnel barrier ($E_{\mathrm{g}}$) is effectively reduced by $\Delta$ due to the energetic separation of the tail states from the ideal VB, (ii) the imaginary dispersion is determined by the one-band effective mass $m_{\mathrm{c}}$ instead of the reduced effective mass $m_{\mathrm{r}}$ due to the assumption of strong localization of the deep tail states, and (iii) the pre-exponential factor is reduced by $\frac{\eta^{3/2}}{\Delta^{3/2}[(\Delta/\eta)^{3/2}+s]}\exp\left(-\frac{\Delta^2}{\eta^2}\right)$. In the low-field range, which is relevant for the slope of a TFET, $(\Delta/\eta)^{3/2} \gg s$ and only *deep* tail states contribute to the generation rate. For an estimate of the ratio $r = G_{\mathrm{tc}}/G_{\mathrm{BTB}}$, the following assumptions for the parameters are made: $m_{\mathrm{r}} = m_{\mathrm{c}}/2, \eta = \hbar\theta_{\mathrm{c}} = E_{\mathrm{g}}/10$. This gives $r = 1.43\times10^{-4}$. Thus, the contribution of *deep* tail states to the total band-to-band generation is negligible.

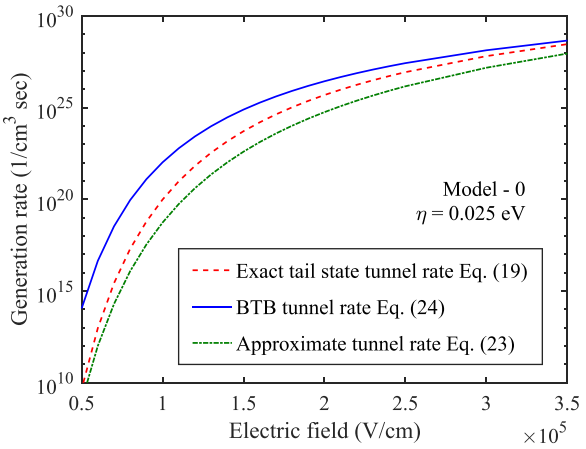


FIG. 4. Generation rates due to strongly localized tail states without field broadening calculated with exact Gaussian tail DOS [Eq. (19)] in comparison to the analytical approximation [Eq. (23)] and the BTBT rate [Eq. (24)].

Figure 4 compares the full model Eq. (19) with the analytical approximation Eq. (23) and the BTBT rate [Eq. (24)] as a function of electric field. The band gap and effective mass of bulk InAs have been used. In the low-field range, the contribution of tail states to the total band-to-band generation is negligible if the life time broadening of their energy levels is ignored. In this range, the analytical expression agrees well with the full model as deep tail states are dominant. With increasing field, shallower states become more and more important and the saddle-point method based on deep Gaussian tails fails at the band edge, causing a growing deviation from the full model.

### C. Tunneling from/into shallow tail states including their life time broadening

The life time broadening is approximately given by $\Delta\tau = \theta_t^{-1}$ which is a measure of the tunnel probability out of the localized tail state. At $F = 1 \times 10^5$ V/cm, the characteristic energy $\hbar\theta_t$ ranges from 20 meV to 12 meV for $m_t$ between 0.5 $m_0$ and 2 $m_0$. This is of the same order as the energies $E_t$ in the tail. Thus, the zero-field DOS of the sharp single-level, Eq. (10), is to be replaced by a properly broadened delta-function. This can be done thanks to an analytical solution of the Schrödinger equation for an effective potential which is the sum of the pseudo-$\delta$-potential and the electrostatic potential in a constant electric field.[1] The DOS Eq. (10) is then replaced by (see Appendix A, Eqs. (A4)–(A9) and also Ref. 17)

$$\varrho_t(\tilde{E},E_t) = \frac{1}{4\pi^3 r_0^3\sqrt{E_t\hbar\theta_t}} \frac{\mathcal{F}\left(\frac{\tilde{E}}{\hbar\theta_t}\right)}{\mathcal{F}^2\left(\frac{\tilde{E}}{\hbar\theta_t}\right) + \left[\mathcal{G}\left(\frac{\tilde{E}}{\hbar\theta_t}\right) + \frac{1}{\pi}\sqrt{\frac{E_t}{\hbar\theta_t}}\right]^2}, \tag{25}$$

which is a Lorentzian-like function of $E_t$ with the property that it approaches the delta-function Eq. (10) for $\hbar\theta_t \to 0$. The peak position is determined by the zero of $\mathcal{G}(\frac{\tilde{E}}{\hbar\theta_t}) + \frac{1}{\pi}\sqrt{\frac{E_t}{\hbar\theta_t}}$. The function $\mathcal{G}(\frac{\tilde{E}}{\hbar\theta_t})$ is given by

$$\mathcal{G}(x) = Ai'(x)Bi'(x) - xAi(x)Bi(x) \to -\frac{\sqrt{x}}{\pi}\left(1 - \frac{0.03123}{x^3}\right) \text{ for large } x > 0 \tag{26}$$

where $Bi(x)$ denotes the Airy function of the second kind. The second term in braces is proportional to $F^2$ and results in a shift of the resonance peak to larger energies (quadratic Stark effect). The peak height is determined by the inverse of the function $\mathcal{F}(\frac{\tilde{E}}{\hbar\theta_t})$ given in Eq. (17).

The transition probability from a field-broadened localized tail state to a field-dependent state in the opposite band is derived in Appendix A. A closed-form expression can be obtained for the cases of strong and very weak localization, respectively. No analytical solution is possible in the general case. The critical parameter is $a^3 = m_c/(m_c + m_t) = \mu/m_t$, where $\mu$ denotes the reduced effective mass $m_c m_t/(m_c + m_t)$. Strong localization is defined by the condition $a^3 \ll 1$, whereas very weak localization by $a^3 \approx 1$.

#### 1. Strong localization

In the case of strong localization ($a^3 \ll 1$), one obtains (see Appendix A)

$$D_{t,c}(E,E',E_t) = 8\pi\frac{m_c}{\mu}\frac{\mathcal{P}}{\tilde{E}^2}E_t^2 r_0^3\,\varrho_t(\tilde{E},E_t)\,\varrho_\mu(E'). \tag{27}$$

Obviously, Eq. (9) is the special case in which the field-effect on the localized states is neglected. The essential difference to Eq. (9) is that the CB DOS $\varrho_c(E')$ given by Eq. (17) is replaced by the *joint* DOS

$$\varrho_\mu(E') = \frac{\sqrt{8\mu^3}}{4\pi\hbar^3}\sqrt{\hbar\theta_\mu}\,\mathcal{F}\left(-\frac{E'}{\hbar\theta_\mu}\right) \tag{28}$$

which contains the reduced effective mass $\mu$ instead of the CB mass $m_c$. This leads to an increased tunnel probability. Note that this result is non-trivial. The occurrence of a reduced effective mass for a transition from a localized state into a Bloch state is due to the assumption that the localized state is built from a single, parabolic band with "effective" mass $m_t$. Note also that $\varrho_t(\tilde{E},E_t)$ is exactly the same as defined in Eq. (25).

With (25), the generation rate is given by

$$G_{tc} = \frac{(eF)^2\sqrt{\eta}}{2^{3/2}\pi^4\hbar^2 E_g}\frac{m_c\sqrt{\mu}}{m_r}\sqrt{\frac{\hbar\theta_\mu}{\hbar\theta_t}}\int_{E_{edge}}^{E_g} dE_t\, Y\left(-\frac{E_t}{\eta}\right) \times \int_{E_{edge}}^{E_g}\frac{d\tilde{E}}{\tilde{E}^2}\frac{\mathcal{F}\left(\frac{E_g-\tilde{E}}{\hbar\theta_\mu}\right)\mathcal{F}\left(\frac{\tilde{E}}{\hbar\theta_t}\right)}{\mathcal{F}^2\left(\frac{\tilde{E}}{\hbar\theta_t}\right) + \left[\mathcal{G}\left(\frac{\tilde{E}}{\hbar\theta_t}\right) + \frac{1}{\pi}\sqrt{\frac{E_t}{\hbar\theta_t}}\right]^2} \tag{29}$$

which after inserting the function $Y$ for Gaussian tails becomes the triple integral

$$G_{\rm tc} = \frac{(eF)^2}{2^{3/2}\pi^{9/2}\hbar^2 E_{\rm g}} \frac{m_{\rm c}\sqrt{\mu}}{m_{\rm r}} \sqrt{\frac{\hbar\theta_\mu}{\hbar\theta_{\rm t}}} \frac{1}{\eta} \times \int_0^\infty d\zeta \sqrt{\zeta} \int_{E_{\rm edge}}^{E_{\rm g}} \frac{d\tilde{E}}{\tilde{E}^2} \mathcal{F}\left(\frac{E_{\rm g}-\tilde{E}}{\hbar\theta_\mu}\right) \mathcal{F}\left(\frac{\tilde{E}}{\hbar\theta_{\rm t}}\right) \times \int_{E_{\rm edge}}^{E_{\rm g}} dE_{\rm t} \frac{\exp\left[-\frac{(\zeta+E_{\rm t})^2}{\eta^2}\right]}{\mathcal{F}^2\left(\frac{\tilde{E}}{\hbar\theta_{\rm t}}\right) + \left[\mathcal{G}\left(\frac{\tilde{E}}{\hbar\theta_{\rm t}}\right) + \frac{1}{\pi}\sqrt{\frac{E_{\rm t}}{\hbar\theta_{\rm t}}}\right]^2} . \quad (30)$$

For an analytical treatment, one can use the fact that even if $\hbar\theta_{\rm t} \approx \eta$, the Lorentzian describing the trap DOS has a sharp maximum at the resonance energy defined by $\mathcal{G}(\frac{\tilde{E}}{\hbar\theta_{\rm t}}) + \frac{1}{\pi}\sqrt{\frac{E_{\rm t}}{\hbar\theta_{\rm t}}} = 0$. At $F = 1\times10^5$ V/cm, a mass as small as $m_{\rm t} = 0.003\, m_0$ results in a broadening of only $\sim$1 meV. Hence, one can replace $\mathcal{F}/(\mathcal{F}^2+\hat{\mathcal{G}}^2) \to 2\pi^2 \sqrt{E_{\rm t}\hbar\theta_{\rm t}}\, \delta(\tilde{E}-E_{\rm t})$. With the asymptotic limit for $\mathcal{F}(\frac{E_{\rm g}-\tilde{E}}{\hbar\theta_\mu})$, this leads to

$$G_{\rm tc} = \frac{(eF)^3}{16\pi^3\hbar E_{\rm g}^2} \frac{m_{\rm c}}{m_{\rm r}} \int_{E_{\rm edge}/\eta}^{E_{\rm g}/\eta} \frac{d\epsilon}{\epsilon^{3/2}} \exp\left[-\frac{4}{3}\left(\frac{E_{\rm g}-\epsilon\eta}{\hbar\theta_\mu}\right)^{3/2}\right] \times \frac{1}{\sqrt{\pi}} \int_0^\infty dt\sqrt{t}\exp\left[-(t+\epsilon)^2\right] . \quad (31)$$

Note that the last line is just $Y_{\rm g}(-\epsilon)$. In the low-field range, only deeper tail states contribute and the saddle-point method can be applied as before. Using approximation (5) for the Gaussian tail DOS yields

$$G_{\rm tc} = \frac{(eF)^3\,\eta^{3/2}}{64\sqrt{2}\pi^{\frac{5}{2}}\hbar E_{\rm g}^2} \frac{m_{\rm c}}{m_{\rm r}} \left(\frac{\mu}{m_{\rm c}}\right)^{\frac{1}{3}} \times \frac{\exp\left[-\frac{\Delta^2}{\eta^2} - \frac{4}{3}\left(\frac{E_{\rm g}-\Delta}{\hbar\theta_\mu}\right)^{\frac{3}{2}}\right]}{\Delta^{3/2}\left[(\Delta/\eta)^{3/2}+s\right]} \quad (32)$$

where $\theta_{\rm c}$ has to be replaced by $\theta_\mu$ in the expression (22) for $\Delta$. Hence, basically Eq. (23) is recovered, with the reduced effective mass $\mu$ instead of the CB effective mass. At high fields (large $\hbar\theta_\mu$) or for small $\eta$, the most contributing energies are close to the VB edge. The factor of the integrand in the first line of Eq. (31) is a relatively smooth function here and can be taken out at the characteristic tail energy $\eta$ ($\epsilon = 1$) or at some fraction $\epsilon^* = E^*/\eta$ of it ($E_{\rm edge}/\eta < \epsilon^* < 1$). The integration limits of the $\epsilon$-integral can be approximately changed such that the remaining double integral becomes

$$\int_0^\infty d\epsilon \int_0^\infty dt\sqrt{t}\exp\left[-(t+\epsilon)^2\right] = \gamma = 0.302 . \quad (33)$$

The error compared to Eq. (31) vanishes with the ratio $E_{\rm edge}/\eta$. The result is

$$G_{\rm tc} = \frac{\gamma\,(eF)^3}{16\pi^{7/2}\hbar E_{\rm g}^2} \frac{m_{\rm c}}{m_{\rm r}} \left(\frac{\eta}{E^*}\right)^{\frac{3}{2}} \exp\left[-\frac{4}{3}\left(\frac{E_{\rm g}-E^*}{\hbar\theta_\mu}\right)^{\frac{3}{2}}\right] . \quad (34)$$

In the case of exponential tails, the factor $\gamma$ has the value 0.785.

The generation rate obtained for the field-broadened tail states [Eq. (29)] is compared with the analytical approximation [Eq. (32)] and the rate of BTBT for reference in Figs. 5 and 6. If the effective mass of the localized state $m_{\rm t}$ is set to the light-hole mass, the generation rate due to tail states almost coincides with the BTBT rate. For strong localization ($m_{\rm t} = 0.41\, m_0$), the effect of field broadening almost disappears and model-1 gives a similar curve to model-0.

### 2. Very weak localization

In the case of very weak localization ($a^3 \approx 1$), one obtains for the transition rate (see Appendix A)

$$D_{\rm t,c}(E,E',E_{\rm t}) = 8\pi \frac{Bi^2\left(\frac{\tilde{E}}{\hbar\theta_\mu}\right)}{(\hbar\theta_\mu)^2} \frac{m_{\rm t}}{\mu} \times E_{\rm t}^2 r_0^3\, \varrho_{\rm t}(\tilde{E},E_{\rm t})\, \varrho_\mu(-E_{\rm g}) \quad (35)$$

which holds for $\hbar\theta_\mu \to \hbar\theta_{\rm t}$.

Instead of Eq. (29), the generation rate is now given by

$$G_{\rm tc} = \frac{(eF)^2\sqrt{\eta}}{2^{3/2}\pi^4\hbar^2 E_{\rm g}} \frac{m_{\rm t}\sqrt{\mu}}{m_{\rm r}} \frac{1}{\sqrt{\hbar\theta_{\rm t}}(\hbar\theta_{\rm t})^{3/2}} \times \mathcal{F}\left(\frac{E_{\rm g}}{\hbar\theta_\mu}\right) \int_{E_{\rm edge}}^{E_{\rm g}} dE_{\rm t} Y\left(-\frac{E_{\rm t}}{\eta}\right) \times \int_{E_{\rm edge}}^{E_{\rm g}} d\tilde{E} \frac{Bi^2\left(\frac{\tilde{E}}{\hbar\theta_\mu}\right)\mathcal{F}\left(\frac{\tilde{E}}{\hbar\theta_{\rm t}}\right)}{\mathcal{F}^2\left(\frac{\tilde{E}}{\hbar\theta_{\rm t}}\right) + \left[\mathcal{G}\left(\frac{\tilde{E}}{\hbar\theta_{\rm t}}\right) + \frac{1}{\pi}\sqrt{\frac{E_{\rm t}}{\hbar\theta_{\rm t}}}\right]^2} . \quad (36)$$

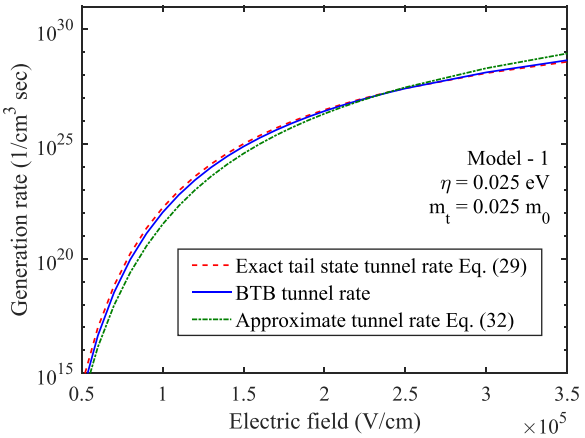


FIG. 5. Generation rates due to strongly localized tail states with $m_{\rm t} = 0.025\, m_0$ including their field broadening calculated with the full model [Eq. (29)] in comparison to the analytical approximation [Eq. (32)] and the BTBT rate [Eq. (24)].

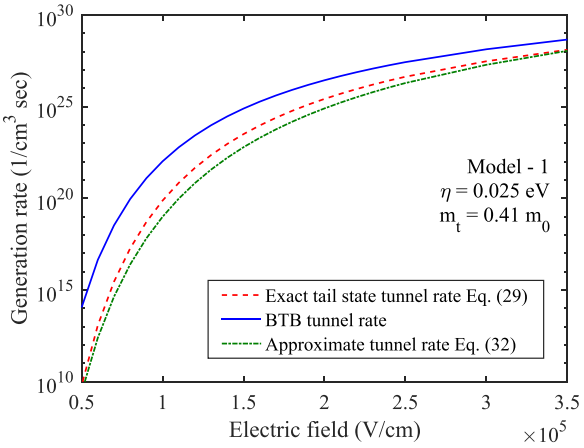


FIG. 6. Generation rates due to strongly localized tail states with $m_t = 0.41\, m_0$ including their field broadening calculated with the full model [Eq. (29)] in comparison to the analytical approximation [Eq. (32)] and the BTBT rate [Eq. (24)].

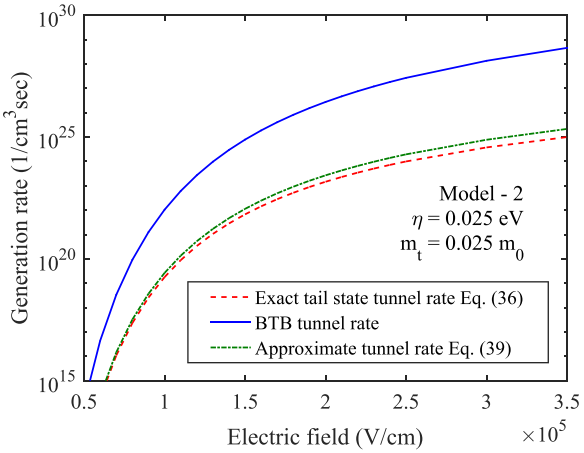


FIG. 7. Generation rates due to very weakly localized tail states with $m_t = 0.025\, m_0$ including their field broadening calculated with the full model [Eq. (36)] in comparison to the analytical approximation [Eq. (39)] and the BTBT rate [Eq. (24)].

As discussed in Subsection II C I, one can apply the replacement $\mathcal{F}/(\mathcal{F}^2+\hat{\mathcal{G}}^2) \to 2\pi^2\sqrt{E_t\hbar\theta_t}\,\delta(\tilde{E}-E_t)$. This simplifies (36) in the case of Gaussian band tails to

$$G_{tc} \approx \frac{\eta^2(eF)^2}{(2\pi)^{7/2}\hbar^2 E_g^2}\frac{m_t\sqrt{\mu}}{m_r}\frac{1}{\sqrt{\hbar\theta_\mu}} \times \exp\left[-\frac{4}{3}\left(\frac{E_g}{\hbar\theta_\mu}\right)^{3/2}\right] \times \int_{E_{edge}/\eta}^{E_g/\eta} d\epsilon \int_0^\infty dt\sqrt{t}\,Bi^2\left(\epsilon\frac{\eta}{\hbar\theta_\mu}\right)\exp\left[-(t+\epsilon)^2\right]. \quad (37)$$

The initial assumption of small masses $m_t$ implies that $\hbar\theta_\mu \to \hbar\theta_t \gg \eta$ and the argument of $Bi^2$ varies slowly. Taking $Bi^2$ out of the $\epsilon$-integral at $\epsilon = E_{edge}/\eta$ then gives

$$G_{tc} \approx \frac{\eta^2(eF)^2}{(2\pi)^{7/2}\hbar^2 E_g^2}\frac{m_t\sqrt{\mu}}{m_r}\frac{Bi^2\left(\frac{E_{edge}}{\hbar\theta_\mu}\right)}{\sqrt{\hbar\theta_\mu}}\exp\left[-\frac{4}{3}\left(\frac{E_g}{\hbar\theta_\mu}\right)^{\frac{3}{2}}\right] \times \int_{E_{edge}/\eta}^{E_g/\eta} d\epsilon \int_0^\infty dt\sqrt{t}\exp\left[-(t+\epsilon)^2\right]. \quad (38)$$

The remaining double integral is the same as in Subsection II C I and can be calculated as described there. The final result is

$$G_{tc} \approx \frac{\gamma\,\eta^2(eF)^2}{(2\pi)^{7/2}\hbar^2 E_g^2}\frac{m_t\sqrt{\mu}}{m_r}\frac{Bi^2\left(\frac{E_{edge}}{\hbar\theta_\mu}\right)}{\sqrt{\hbar\theta_\mu}} \times \exp\left[-\frac{4}{3}\left(\frac{E_g}{\hbar\theta_\mu}\right)^{\frac{3}{2}}\right]. \quad (39)$$

The generation rate obtained for the full model [Eq. (36)] is compared with the analytical approximation [Eq. (39)] and the rate of BTBT for reference in Fig. 7. If the effective mass of the localized state $m_t$ is set to the light hole mass, the full model and analytical approximation agree very well. Decreasing $m_t$ shifts the curves up, but a value as small as $0.01\ m_0$ is needed to match the BTBT curve at low fields. This demonstrates that model-2 is completely inappropriate. It is evident that tail-to-band tunneling becomes comparable to BTBT for some field value in the sub-threshold regime. For $m_t = m_c$, the rate of tail-to-band tunneling must be in the same order of magnitude as the rate of BTBT. This is the outcome in the case of strong localization (model-1, see Fig. 5) which is, therefore, the recommended model.

### D. Tunneling from/into continuum states of the non-ideal DOS

States with energies $\tilde{E} < E_{edge}$ are considered as Bloch states. However, compared to the usual description of direct BTBT [e.g., Eq. (24)] the rate will be different due to the modified shape of the DOS at these energies. The goal of this subsection is to find a proper replacement for Eq. (24) if the VB DOS is non-ideal.

In a first step, the ideal BTBT rate, which is proportional to the reduced DOS, is written as convolution of the ideal CB and VB DOSs:

$$G_{BTB} = \frac{(eF)^3}{8\hbar E_g\hbar\theta_r}\mathcal{F}\left(\frac{E_g}{\hbar\theta_r}\right) = c\int_{-\infty}^{\infty} dE\,\varrho_v(\tilde{E})\,\varrho_c(E) = c\,\frac{(2m_v)^{3/2}(2m_c)^{3/2}}{(2\pi\hbar^3)^2}\sqrt{\hbar\theta_v}\sqrt{\hbar\theta_c} \times \int_{-\infty}^{\infty} dE\,\mathcal{F}\left(\frac{\tilde{E}}{\hbar\theta_v}\right)\mathcal{F}\left(\frac{-E}{\hbar\theta_c}\right) \quad (40)$$

[see Eq. (17)]. In Appendix B, it is demonstrated that the integral in Eq. (40) can be calculated exactly, but $c$ becomes a complicated function of various Airy functions then. A simple form of $c$ turns out in the asymptotic (WKB) limit which holds if $E_g \gg \hbar\theta_r$:

$$c = \frac{\pi^{7/2}}{\hbar}\left(\frac{\hbar^2}{m_\Sigma}\right)^{3/2}\frac{(\hbar\theta_r)^{3/4}}{E_g^{1/4}}. \tag{41}$$

In a second step, the ideal field-broadened VB DOS $\varrho_v(\tilde{E})$ is replaced by the non-ideal field-broadened VB DOS $\varrho_{v,t}(\tilde{E})$,

$$\varrho_{v,t}(\tilde{E}) = \left(\frac{2m_v}{\hbar^2}\right)^{3/2}\frac{\sqrt{\eta}}{2\pi\sqrt{\hbar\theta_v}}\int_{-\infty}^{\infty} d\epsilon\,\Theta(\epsilon)\frac{1}{\sqrt{\epsilon}} \times Y\left(\frac{\epsilon - E_{edge}}{\eta}\right)Ai^2\left(\frac{\epsilon - E_{edge} + \tilde{E}}{\hbar\theta_v}\right). \tag{42}$$

Note that $\varrho_v(\tilde{E})$ is reproduced in the limit $\eta \to 0$ since $E_{edge}$ is always a fraction of $\eta$. The proof of Eq. (42) is presented in Appendix C. The calculation of the rate of transitions from continuum states of the non-ideal VB DOS into CB states now proceeds in a similar way as outlined in Appendix B for the ideal VB DOS. Appendix D contains details of this derivation. Inserting $c$, it follows exactly:

$$G_{tc}^{cond} = \frac{(eF)^3}{2\hbar E_g}\left(\frac{E_g}{\hbar\theta_r}\right)^{\frac{3}{4}}\frac{\sqrt{\pi}\sqrt{\eta}}{(\hbar\theta_r)^2}\int_0^{\infty}\frac{d\epsilon}{\sqrt{\epsilon}}Y\left(\frac{\epsilon - E_{edge}}{\eta}\right) \times \int_0^{\infty} dt\,\sqrt{t}Ai^2\left(t + \frac{\epsilon + E_g'}{\hbar\theta_r}\right) \tag{43}$$

where $E_g' = E_g - E_{edge}$. The last step is to convert the double integral into two useful forms: (i) a convolution integral of the ideal field-broadened BTBT rate $G_{BTB}$ with a "smoothing function" $S$, and (ii) a fully analytical expression to better understand the main effects that the non-ideal DOS has on the rate. As shown in Appendix D, the first form can be written as

$$G_{tc}^{cond} = \int_0^{\infty} d\epsilon\, S(\epsilon)\, G_{BTB}\left(\frac{\epsilon + E_g'}{\hbar\theta_r}\right) \tag{44}$$

with

$$S(\epsilon) = \left(\frac{E_g}{\hbar\theta_r}\right)^{\frac{1}{2}}\frac{\sqrt{2\pi}}{\hbar\theta_r}\sqrt{\frac{\eta}{\epsilon}}Y\left(\frac{\epsilon - E_{edge}}{\eta}\right). \tag{45}$$

The fully analytical form, valid in the asymptotic limit, reads (see Appendix D)

$$G_{tc}^{cond} = \frac{(eF)^3}{64\pi\hbar E_g^2}\frac{2\sqrt{\eta}E_g^{1/4}}{(\hbar\theta_r)^{3/4}} \times Y\left[\frac{(\hbar\theta_r)^{3/2}}{4\eta E_g^{1/2}} - \frac{E_{edge}}{\eta}\right]e^{-\frac{4}{3}\left(\frac{E_g'}{\hbar\theta_r}\right)^{\frac{3}{2}}}. \tag{46}$$

As shown in Appendix D, the pre-factor has been slightly adjusted to ensure that $\lim_{\eta\to 0} G_{tc}^{cond} = G_{BTB}$. Equation (46) reveals two non-ideality effects of the DOS: (i) an effective reduction of the gap by $E_{edge}$ and (ii) a smoothing effect described by a factor which depends on the four characteristic energies $E_g$, $\hbar\theta_r$, $\eta$, $E_{edge}$, and the tail shape function $Y$. There is no analytical approximation for the smoothing factor $\sqrt{z}\,Y(1/z - \xi) \overset{\eta\to 0}{\to} 1$ around the band edge. The factor is plotted as function of $\eta$ for different values of the ratio $E_{edge}/\eta$ in Fig. 8. The plots are normalized by the generation rate at $\eta = 0$. The rate first decreases with increasing $\eta$ due to the redistribution of VB states which leads to a reduction of the VB DOS at the band edge. As $\eta$ increases further, the shrinking effective tunnel gap increases the tunnel rate. However, the overall reduction of the generation rate is very weak. Even for a ratio $E_{edge}/\eta$ as large as 0.55, the maximum reduction of the rate amounts to only ≈25%. Thus, one can conclude that the modification of the rate due to the redistribution of extended states of the VB DOS is a second-order effect and can be ignored.

### E. Comparison of the localization regimes

The tail-to-band generation rates in a constant electric field were evaluated numerically for three cases (see Table I): Eq. (19) (strong localization without field broadening → model-0), Eq. (29) (strong localization with field broadening → model-1), and Eq. (36) (weak localization with field broadening → model-2). They are shown as a function of electric field in Fig. 9. The band gap and effective masses of bulk InAs have been used for the calculation. The effective mass $m_t$ was set to a value of $0.025\,m_0$ (i.e., close to the measured light-hole mass in InAs) for calculating the generation rate in the cases of model-1 and model-2 (note that model-0 is independent of $m_t$). This choice agrees with the masses to be used in BTBT, and consequently, the rate from model-1 is very close to the BTBT rate. Model-0 corresponds to the limit $m_t \to \infty$ in model-1, and results in a rate which is four orders of magnitude smaller at low fields. This demonstrates the importance of lifetime broadening of the localized state which must be taken into account. The generation rate calculated with model-2 is small compared to that from model-1 which just indicates the strong violation of the condition $m_t \ll m_c$.

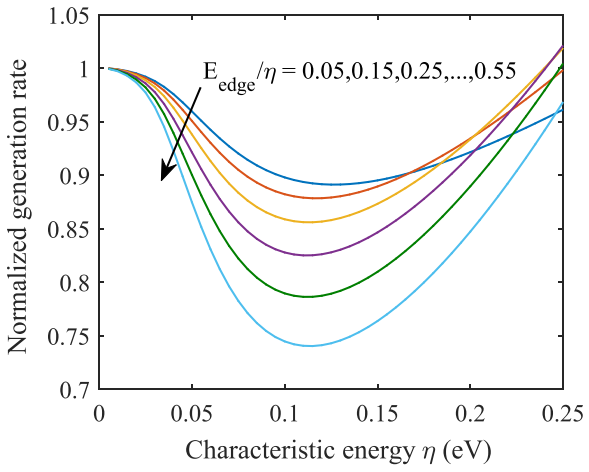


FIG. 8. Generation rate between extended states of the VB DOS and the CB calculated using Eq. (46) versus characteristic energy $\eta$ at different values of $\frac{E_{edge}}{\eta}$. The electric field was set to 1 MV/cm.

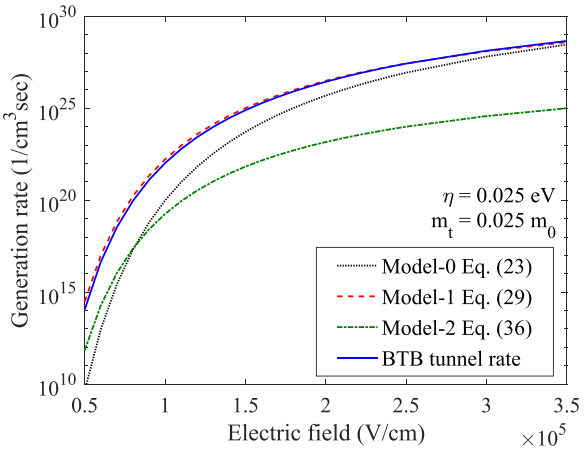


FIG. 9. Comparison of different models for tunneling from DOS tail states into the CB. The BTBT rate [Eq. (24)] based on the same dispersion is also shown. Curves correspond to the model of strong localization without field broadening [model-0—Eq. (19)], with field broadening [model-1—Eq. (29)], and the model of weak localization [model-2—Eq. (36)].

## III. DEVICE APPLICATION

### A. Formulation for the Dynamic Nonlocal Path (DNLP) algorithm

The analytical forms of the generation rates for tunneling between VB tail states and CB states have been derived for a long semiconductor region with constant electric field. Using them as local models in TCAD would result in an over-estimation of the generation current in TFETs. This is illustrated in Fig. 10. A tunnel path starting from the CB edge at $x = x_1$ ends at the VB edge. Therefore, electrons from all tail states throughout the band gap ($E_{\text{edge}} < E_t < E_g$) tunnel to the CB at $x_1$. However, a tunnel path starting at $x = x_2$ on the CB edge can only be used by tail states with an energy above the mid-gap ($E_t > E_g/2$). As a result, the tail states above the mid-gap ($E_{\text{edge}} < E_t < E_g/2$) are active. A calculation of the electron generation at $x_1$ using the expressions in Sec. II would yield correct results. However, employing the same expressions to calculate the generation rate at $x_2$ would over-estimate the tunnel current. Therefore, the numerical implementation must be generalized to inhomogeneous electric fields. This is done in the commercial device simulator Sentaurus-Device[28] by the so-called *Dynamic Nonlocal Path (DNLP) Algorithm* which calculates the tunnel rate by numerical integration over the

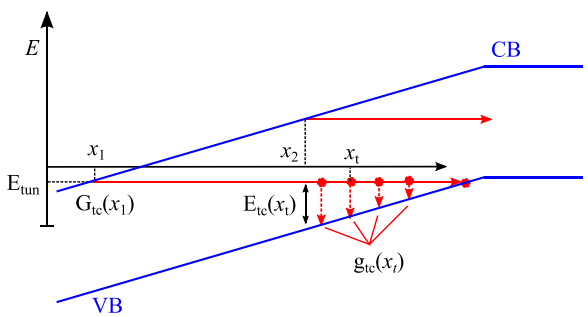


FIG. 10. Schematic representation of two generic types of tunnel paths occurring in a TFET. The path beginning at $x_2$ does not come across tail states with $E_t < E_{\text{mid}}$. Various variables used in Eqs. (48)–(51) are shown.

action along dynamically extracted tunnel paths. The algorithm checks if a tunnel path is actually active. Thus, the expressions derived in Sec. II need to be adapted to the DNLP algorithm.

In general, all energetic tunnel rates are transformed into position-dependent rates defined along a tunnel path. The adaptation of Eq. (19) proceeds as follows. Each tunnel path connects the CB edge and VB edge. At any point along the tunnel path located at a distance $x_t \in \{x_1, x_1 + L\}$, $\tilde{E} \to E_{\text{tun}} - E_v(x_t)$ and $d\tilde{E} \to eF(x_t)\Delta x_t$. In this way, Eq. (19) becomes locally defined at $x_1$:

$$G_{tc}(x_1) = \sum_{x_t = x_1}^{x_t = x_1 + L} g_{tc}(x_t) \tag{47}$$

$$g_{tc}(x_t) = \frac{(eF_{av}(x_t))^2 \sqrt{\eta}\sqrt{\hbar\theta_c} m_c^{\frac{3}{2}}}{\sqrt{2}\pi^{5/2}\hbar E_g^2 m_r} \frac{eF(x_t)\Delta x_t}{(E_{tun} - E_v(x_t))^{\frac{3}{2}}} \times Y\left(-\frac{E_{tun} - E_v(x_t)}{\eta}\right) \mathcal{F}\left(\frac{E_c(x_t) - E_{tun}}{\hbar\theta_c}\right) \times \left[f_n(x_1) - f_p(x_t)\right]. \tag{48}$$

Here, $L$ is the length of the tunnel path, $x_t \in \{x_1, x_1 + L\}$ is the location of the tail state along the tunnel path, $\Delta x_t$ is the discretization interval, $F_{av}(x_t) = \frac{1}{|x_t - x_1|}\int_{x_1}^{x_t} F(x)dx$ is the electric field averaged over the segment of the tunnel path between $x = x_1$ and $x = x_t$, and $E_{tun}$ is the CB energy at the beginning of the tunnel path. $E_c(x_t)$, $E_v(x_t)$, and $F(x_t)$ are, respectively, the CB edge, the VB edge, and the electric field at the location of the tail state, $g_{tc}$ is the generation rate at $x_t$, and $G_{tc}$ is the total generation rate at $x_1$. The function $f_{n/p}(x) = [\exp(E_{tun} - E_{F,n/p}(x))/k_B T + 1]^{-1}$ represents the Fermi distribution at $x$. The value of $Y(-\frac{E_{tun} - E_v(x_t)}{\eta})$ is calculated at each $x_t$ using Eq. (2) or Eq. (8) for Gaussian or exponential tails, respectively.

If field broadening of the tail states is considered, the generation rate due to tail states is given by Eq. (29) ($a^3 = \frac{m_c}{m_c + m_t} \ll 1$) which is modified by making the substitutions $E_t \to E_{tun} - E_v(x_t) = E_t(x_t)$ and $dE_t \to eF(x_t)\Delta x_t$. If $\tilde{E}$ in the second integral is replaced by $\epsilon\,\hbar\theta_t$, the generation rate $g_{tc}(x_t)$ becomes

$$g_{tc}(x_t) = \frac{(eF_{av}(x_t))^2 \sqrt{\eta}}{2^{3/2}\pi^{9/2}\hbar^2 E_g} \frac{m_c\sqrt{\mu}}{m_r\hbar\theta_t} \sqrt{\frac{\theta_\mu}{\theta_t}} eF(x_t)\Delta x_t \times Y\left(-\frac{E_t(x_t)}{\eta}\right) \mathcal{H}\left(\frac{E_g}{\hbar\theta_\mu}, \frac{E_t(x_t)}{\hbar\theta_t}\right) \times \left[f_n(x_1) - f_p(x_t)\right] \tag{49}$$

with

$$\mathcal{H}(p,q) = \int_{E_{\text{edge}}/\hbar\theta_t}^{E_g/\hbar\theta_t} \frac{d\epsilon}{\epsilon^2} \frac{\mathcal{F}(p - a\,\epsilon)\mathcal{F}(\epsilon)}{\mathcal{F}^2(\epsilon) + \left[\mathcal{G}(\epsilon) + \frac{1}{\pi}\sqrt{q}\right]^2}. \tag{50}$$

The total generation rate at $x_1$ is calculated by the sum (47).

In the same manner, the expression for tail-to-band tunneling in the case of very weak localization [Eq. (36)] can be modified for TCAD simulations. The role of the function $\mathcal{H}$ is adopted by the function

$$\mathcal{B}(p,q) = \int_{E_{\text{edge}}/\hbar\theta_t}^{E_g/\hbar\theta_t} d\epsilon \frac{Bi^2(a\,\epsilon)\mathcal{F}(\epsilon)}{\mathcal{F}^2(\epsilon) + \left[\mathcal{G}(\epsilon) + \frac{1}{\pi}\sqrt{q}\right]^2}. \quad (51)$$

The functions $Y$, $\mathcal{H}$, and $\mathcal{B}$ can be implemented in the form of look-up tables which are evaluated at the beginning of the simulation run, once the values of $m_c, m_v$, and $m_t$ are known. They are computed at any given input by interpolating between pre-evaluated values using cubic splines. For the integrals to be accurate, the input $q = E_t(x_t)/\hbar\theta_t$ in Eqs. (50) and (51) needs to be less than the upper limit of the integral, i.e., $E_t(x_t) < E_g$. For look-up tables, it is necessary that the integration limits are fixed. For the lower limit, $E_{\text{edge}}/\hbar\theta_t = 0.03$ is used and $E_g/\hbar\theta_t = 10$ for the upper. With the fixed lower limit, all contributions to the integral are safely embraced, and the divergency at $E_{\text{edge}} = 0$ is excluded. In this way, the double/triple integrals in Eqs. (48) and (49) can be transformed into a single integral along the tunnel path. The equations have been implemented in the Finite-Element-based TCAD simulator Sentaurus-Device using the Physical Model Interface (PMI) *Nonlocal Generation-Recombination*. The original DNLP BTBT model requires the effective tunnel barrier and the electron/hole effective masses as input parameters. In addition to these parameters, for the new DNLP tail-to-band tunneling model one has to provide the effective mass $m_t$ of a hole (or electron for the complementary process) localized in the tail state as well as the characteristic energy $\eta$ of the DOS tail.

Note that the implemented tail-to-band tunneling models employ an effective average electric field for the computation of the rate in contrast to the DNLP BTBT model in Sentaurus-Device where the action integral over the imaginary dispersion is computed numerically. The loss in accuracy may be compensated through calibration of the parameters $\eta$ and $m_t$.

### B. Implementation of the DNLP Tail-to-band tunneling model

A semiconductor may exhibit both CB and VB tail states with different characteristic energies $\eta$. This general case is approximated here by the sum of the rates for VB-tail-to-CB tunneling and CB-tail-to-VB tunneling. The expressions for the latter are straightforwardly obtained by obvious changes in the notation of parameters occurring in the former, e.g., $m_c \to m_v, \theta_c \to \theta_v$. The contribution from tail-to-tail tunneling is neglected due to the small probability when both initial and final states are localized. A further refinement of the model would be achieved if in the derivation the ideal DOS of the final states is replaced by the continuum states of the non-ideal DOS, i.e., Eq. (42) for the VB and its counterpart for the CB.

As in the original DNLP model for BTBT, the proposed model involves searching for active tunnel paths. A tunnel path (a straight line in the semi-classical treatment) starting at the VB edge must have an end point *at the CB edge* at the same energy. If no such point is found, the path is discarded. Once all the active tunnel paths at a given bias voltage are found, the tail-to-band tunnel rates are calculated at each discretization point along the tunnel path using one of the three models described above and summed up to obtain the rate at the starting point. Tunneling of an electron from the tail state generates a hole at the same location since it is implicitly assumed that thermionic emission into the VB continuum is very rapid and thus not rate-limiting. The densities of generated holes and electrons enter the Poisson equation and self-consistently impact the solution of the drift-diffusion equation system.

In a TFET, two cases occur that need special treatment. They are sketched in Fig. 11. Figure 11(a) presents the case in which a tunnel path encounters an insulator interface instead of the CB edge. In this situation, the tunnel path is accepted if the energy difference between the tunnel energy ($\epsilon$) and the CB edge at the end point is smaller than a cut-off energy ($E_{\text{cutoff}}$). The integration of the generation rate is performed over all tail states along the path up to the intersection point. The second special case [Fig. 11(b)] occurs when the tunnel path exceeds a maximum length (set to 100 nm) but is still fully contained in the semiconductor. Then the cut-off energy is set to five times the characteristic tail energy ($\eta$) or to half of the gap, whichever is smaller. The choice of the band tail shape (Gaussian or exponential) rests upon the user.

## IV. SIMULATION RESULTS

In order to analyze the impact of band tails on the transfer characteristics of TFETs, a representative sample was simulated with Sentaurus-Device. The structure is a radially symmetric InAs nanowire with a diameter of 10 nm and a gate of 80 nm length. Its radial cross section is shown in Fig. 12(a). The gate is overlapped with the heavily doped source (p-type, $N_A = 1 \times 10^{19}$ cm$^{-3}$) making it a Gate-overlapped-Source (GoS) Nanowire TFET. The electron and light-hole

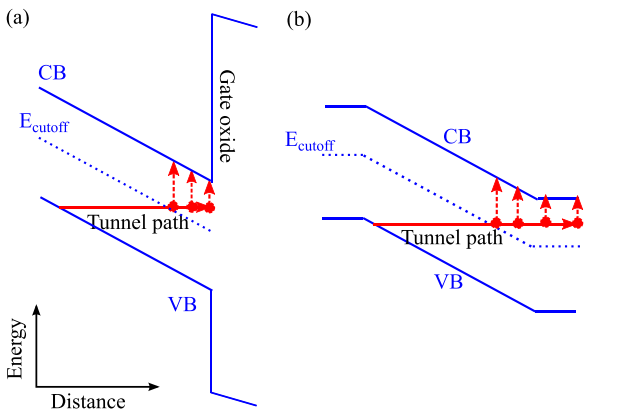


FIG. 11. Treatment of special cases - (a) The tunnel path intersects the semiconductor-oxide interface instead of the CB edge. (b) The tunnel path reaches a maximum length without intersecting the CB edge. In both cases, the tunnel path is accepted if at its end point the energy difference between the CB edge and tunnel energy is smaller than a cut-off energy.

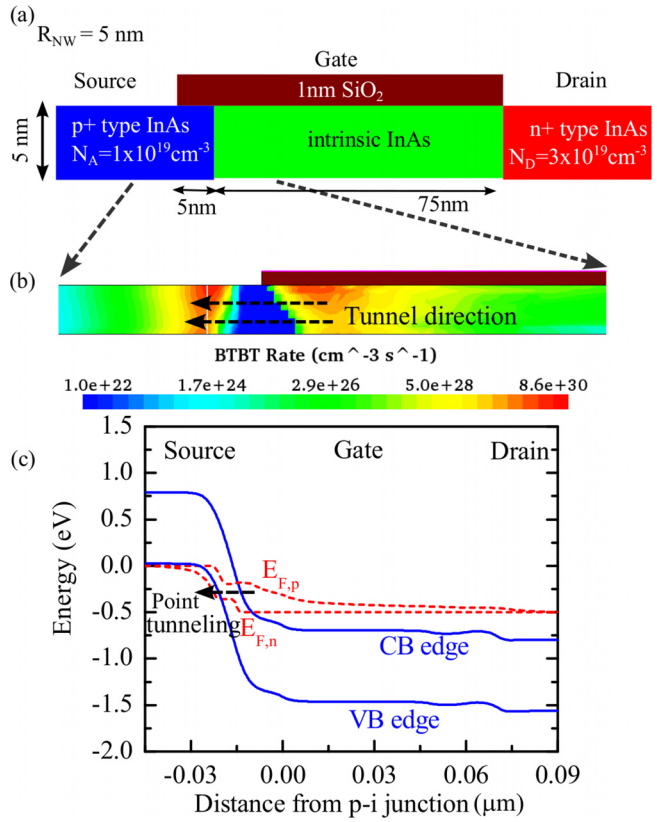


FIG. 12. (a) Radial cross section of the cylindrical nanowire TFET simulated to analyze the impact of tail states. (b) Color-mapped diagram of the BTBT rate in the radial cross section. (c) Band diagram along the axis of the nanowire showing that the tunnel window is opened at the source edge of the gate.

effective mass values of InAs were set to 0.023 $m_0$ and 0.026 $m_0$, respectively.[33] As a result of geometrical confinement, the InAs band gap increases from 0.36 eV to 0.76 eV.[34] Band gap narrowing (BGN) due to heavy doping was neglected in the simulations. The inclusion of BGN would reduce the tunnel gap for both BTBT and tail-to-band tunneling thereby scaling up both rates. To first order, this does not much change the relative strength of both generation mechanisms. BTBT between the CB and the heavy-hole band was also ignored.

The TFET operates as follows. With increasing gate bias, accumulation begins in the channel region. A further increase in the gate voltage pushes the entire band edge down and a tunnel window opens at the source edge of the gate as shown in Fig. 12(c). Electrons tunnel from the source region to the channel along tunnel paths parallel to the gate (so-called point tunneling) as illustrated in Fig. 12(b). Due to the accumulation in the channel region, tunneling normal to the gate (so-called line tunneling) does not take place.

In the following, the effect of various model parameters such as the characteristic energy of the tail $\eta$, the tail shape, and the effective mass $m_t$ on the transfer characteristics will be analyzed. In all cases, model-1 developed above is applied to the device shown in Fig. 12(a).

The characteristic energy $\eta$ of the tail determines how deep the latter penetrates into the gap, although the penetration depth is different for Gaussian and exponential tails. The value of $\eta$ was varied over a feasible range to analyze its impact. The simulated transfer characteristics in the case of exponential tails are presented in Figs. 13(a) and 13(b) for two values of the parameter $m_t$. One can distinguish two distinct branches originating from tail-to-band tunneling and BTBT, respectively. This distinction is more pronounced for smaller $\eta$ or larger $m_t$. As observed, a decreasing $\eta$ steepens the transfer characteristics. For the range of the effective mass $m_t$, an interval between the light-hole mass and the valence band DOS mass was assumed (see the discussion in Subsec. II B). An increasing $m_t$ reduces the drain current arising from tail-to-band tunneling as a consequence of its reduced rate. The almost constant drain current prior to the onset of tunneling arises from SRH generation of electron-hole pairs in the depletion region. Carrier lifetimes of $10^{-9}$ s were used.

The average sub-threshold swing (SS) was calculated for different values of $\eta$ by averaging the inverse slope of the transfer curve over the range from $10^{-15}$ A to $10^{-11}$ A in the drain current. The result is shown in Fig. 14 as a function of $\eta$ for the two values of $m_t$. Increasing $\eta$ degrades the average SS, but the degradation tends to saturate at larger $\eta$. Reducing $m_t$ from 0.41 $m_0$ to 0.025 $m_0$ results in an only small increase of SS. A possible reason for this relative insensitivity of the swing to a variation of $m_t$ could be the following. The steep onset of BTBT in the considered TFET results from the energetic alignment of the DOS in the source with the DOS in the channel at a certain gate bias [see Fig. 1(a)]. The onset is gradual due to the presence of tail states in the heavily doped source [see Fig. 1(b)] which smoothen the band edge. Therefore, the SS depends only on $\eta$ which defines the degree of smearing. The parameter $m_t$ merely determines the tunnel rate between the tail

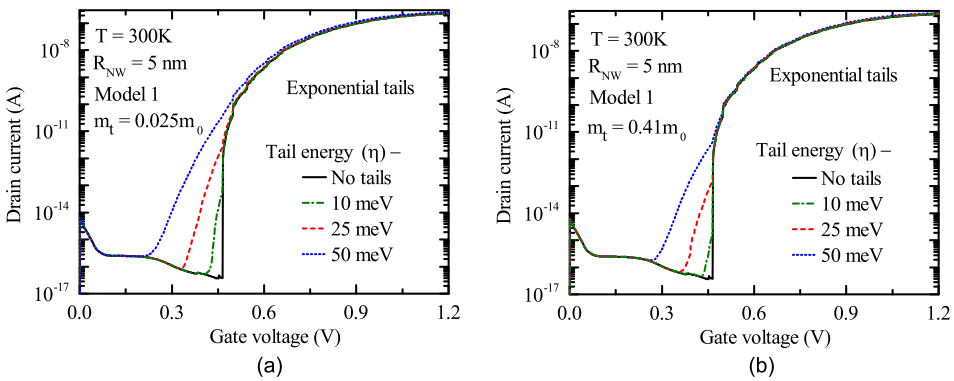


FIG. 13. Impact of the characteristic energy $\eta$ of an exponential DOS tail on the transfer characteristics of the TFET.

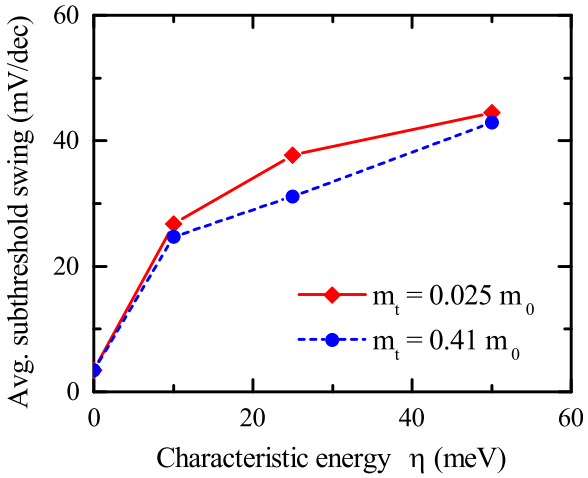


FIG. 14. Degradation of the sub-threshold swing SS with characteristic energy $\eta$ for different values of the effective mass $m_t$.

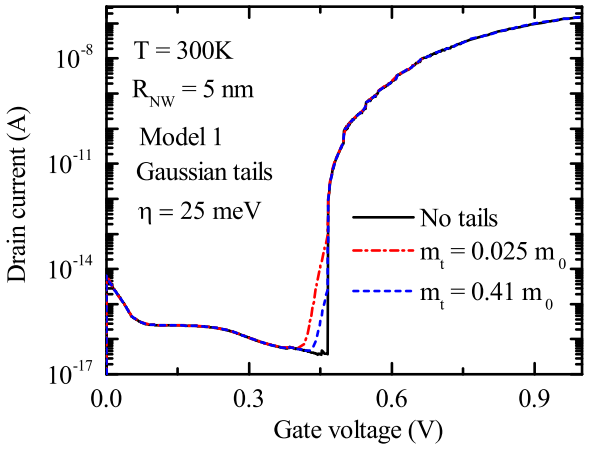


FIG. 16. Impact of the effective mass $m_t$ in the case of a Gaussian DOS. Model-1 which accounts for lifetime broadening of the tail states was used.

state and the CB state. The transfer characteristics for different $m_t$ presented in Fig. 15 show that an increasing $m_t$ scales down the tail-to-band current but does not significantly change the swing.

The transfer characteristics of the TFET for different values of $m_t$ resulting from a Gaussian shape of the DOS tail are plotted in Fig. 16. Again, the drain current is reduced with increasing $m_t$. A comparison of the transfer characteristics of the device with exponential and Gaussian DOS tails, respectively, implies that the exponential shape of the DOS tails degrades the TFET performance more severely than the Gaussian shape. This is due to the fact that, for any given $\eta$, a band edge with Gaussian smoothing is sharper compared to a band edge with exponential smoothing. PL measurements infer the worst case, i.e., the presence of exponential DOS tails.

## V. CONCLUSION

A theory of tail-to-band tunneling in semiconductors has been developed. Compared to prior art, the localized nature of tail states was taken into account. For the complete picture, the field-induced lifetime broadening of these localized states was computed from a 3-dimensional pseudo-delta potential.[1] The basic idea has been to compose the tail DOS as a weighted integral over single-level trap DOSs where the envelope is determined by the measured shape (exponential or Gaussian) and characteristic energy ($\eta$). The deeper the energy in the tail, the stronger the localization and the smaller the probability of trap-assisted tunneling between the tail state and opposite band. The simple potential model yields s-like states with a certain localization radius which is parametrized by an effective mass $m_t$ of the localized electron. This is the only fitting parameter of the theory, but its value can be assumed to be in the range between light-hole and VB DOS (heavy hole) mass. The explicit calculation of the generation rates is only possible in the limits of strong and weak localization, respectively. However, the systematic analysis has shown that the important $m_t$-range is best described by the limit of strong localization. For a homogeneous electric field, the complicated multiple integrals have been simplified to fully analytical expressions with high accuracy and compared to the band-to-band tunneling rate using the same one-band envelope method. An analytical solution for the generation rate due to tunneling from continuum states of the disturbed DOS to states in the opposite band was also derived. The difference to ordinary band-to-band tunneling (Franz-Keldysh effect), however, is small and almost negligible.

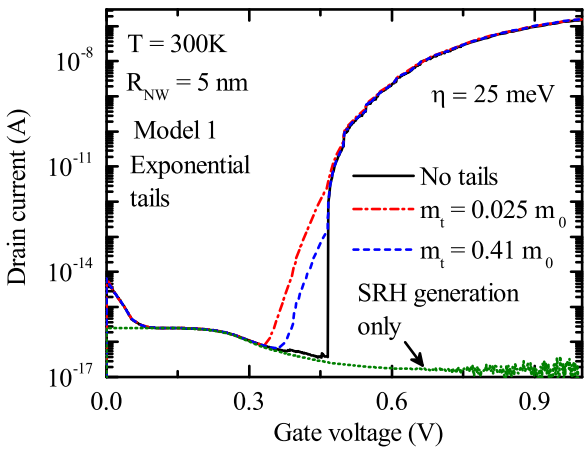


FIG. 15. Comparison of the transfer characteristics obtained using model-1 with different values of the effective mass $m_t$.

In a Tunnel FET, the presence of DOS tails is an important non-ideality effect. The full model was implemented in a commercial semi-classical device simulator and applied to a nanowire Gate-All-Around TFET. As a consequence of the smoothened band edge, the energetic overlap of initial and final states occurs gradually at the onset of tunneling. This increases the sub-threshold swing of the TFET. Exponential tails have a much stronger impact here than Gaussian tails. The localization parameter $m_t$ hardly changes the swing, but determines the magnitude of the generation rate. When tail states are assumed to be Bloch states, the rate is strongly overestimated. Hence, DOS tails degrade the performance of TFETs, but to a lesser extent than trap-assisted tunneling via

deep levels in bulk, at hetero-interfaces, and at oxide-semiconductor interfaces.[5] The characteristic energy $\eta$ has a clear correlation with the doping concentration in the source of the TFET. Simulations have shown that there is an optimum doping level which should not be exceeded since screening then reduces the gate control.[41] This practically limits $\eta$ and the effect of tails on the TFET performance. The developed model is a "continuum model", i.e., it implicitly assumes a proper average over random disorder caused by doping. Aggressive geometrical scaling of nanowire TFETs leads to a countable number of doping atoms in the source region. Then, the occurrence of tails becomes questionable and only an atomistic approach, like tight-binding NEGF, is able to correctly simulate the effect.[12,13]


## ACKNOWLEDGMENTS

This work was supported by the European Communitys Seventh Frame-work Programme under Grant 619509 through Project E2SWITCH.


## APPENDIX A: DERIVATION OF ENERGETIC TAIL-TO-BAND TUNNEL RATES

This appendix outlines the derivation of two approximations for the energetic rate of transitions from localized tail states near the valence band (VB) into conduction band (CB) states [Eqs. (27) and (35)]. Cylinder coordinates are chosen as the constant field $F$ is assumed to be aligned with the $z$-direction. The normalized ground state $\Phi_{\tilde{E}00}$ in the potential $-eFz + 4\pi E_t r_0^3 \delta(\mathbf{r})[1 + \mathbf{r}\cdot\nabla_{\mathbf{r}}]$ is given by[1]

$$\Phi_{\tilde{E}00}(\varrho,z) = \frac{\sqrt{\dfrac{eF\hbar^2}{2\pi(\hbar\theta_t)^3 m_t}}}{\sqrt{\mathcal{F}\left(\dfrac{\tilde{E}}{\hbar\theta_t}\right)}\sqrt{\mathcal{F}^2\left(\dfrac{\tilde{E}}{\hbar\theta_t}\right)+\hat{\mathcal{G}}^2\left(\dfrac{\tilde{E}}{\hbar\theta_t}\right)}} \int_0^\infty d\kappa\,\kappa J_0(\kappa\varrho)\left\{\hat{\mathcal{G}}\left(\frac{\tilde{E}}{\hbar\theta_t}\right)Ai\left(\frac{E_t(\kappa)+\tilde{E}}{\hbar\theta_t}\right)Ai\left(\frac{E_t(\kappa)+\tilde{E}+eFz}{\hbar\theta_t}\right)\right.$$
$$\left. -\mathcal{F}\left(\frac{\tilde{E}}{\hbar\theta_t}\right)\left[\Theta(z)Ai\left(\frac{E_t(\kappa)+\tilde{E}+eFz}{\hbar\theta_t}\right)Bi\left(\frac{E_t(\kappa)+\tilde{E}}{\hbar\theta_t}\right)+\Theta(-z)Ai\left(\frac{E_t(\kappa)+\tilde{E}}{\hbar\theta_t}\right)Bi\left(\frac{E_t(\kappa)+\tilde{E}+eFz}{\hbar\theta_t}\right)\right]\right\} \tag{A1}$$

with

$$\hat{\mathcal{G}}\left(\frac{\tilde{E}}{\hbar\theta_t}\right) = \mathcal{G}\left(\frac{\tilde{E}}{\hbar\theta_t}\right) + \frac{1}{\pi}\sqrt{\frac{E_t}{\hbar\theta_t}}. \tag{A2}$$

In Eq. (A1), $J_0$ denotes the Bessel function of the first kind, of zero order, and $E_t(\kappa) = \hbar^2\kappa^2/(2m_t)$.

In the calculation of the matrix element between the state $\Phi_{\tilde{E}00}$ and the envelope wave function $\Phi_{E'k_\perp m}$ of the CB

$$\Phi_{E'k_\perp m}(\varrho,\varphi,z) = \sqrt{\frac{eFk_\perp}{2\pi}}\frac{1}{\hbar\theta_c}J_m(k_\perp\varrho)e^{im\varphi} \times Ai\left(\frac{E_c(k_\perp)-E'-eFz}{\hbar\theta_c}\right) \tag{A3}$$

the first term in the curly braces in Eq. (A1) can be neglected because $|\mathcal{G}| \ll |\mathcal{F}|$ for the relevant energies $\tilde{E} = E + E_g$. After performing the $\varrho$- and $\varphi$-integration, one obtains

$$M^2_{k_\perp,m}(E,E',E_t) \equiv \left|\left(\Phi_{E'k_\perp m}|\Phi_{\tilde{E}00}\right)\right|^2 = \frac{\delta_{m0}}{(\hbar\theta_t)^3}\frac{\hbar^2 k_\perp}{m_t}\frac{\mathcal{F}\left(\dfrac{\tilde{E}}{\hbar\theta_t}\right)}{\mathcal{F}^2\left(\dfrac{\tilde{E}}{\hbar\theta_t}\right)+\hat{\mathcal{G}}^2\left(\dfrac{\tilde{E}}{\hbar\theta_t}\right)}|C(x,y)|^2 \tag{A4}$$

and

$$C(x,y) = \frac{1}{\pi}\int_{-\infty}^{\infty} dt Ai\left(\frac{\theta_t}{\theta_c}t + x\right) \times \int_{-\infty}^{\infty} d\lambda \frac{\mathcal{P}}{\lambda}Ai(y-t+\lambda)Ai(y+\lambda) \tag{A5}$$

and

$$x = \frac{E_c(k_\perp)-E'}{\hbar\theta_c}, \quad y = \frac{E_t(k_\perp)+\tilde{E}}{\hbar\theta_t}. \tag{A6}$$

The energetic transition rate for one localized state follows from Eq. (A4) by summation over $m$ and integration over $k_\perp$

$$\tilde{D}_{t,c}(E,E',E_t) = \sum_m \int_0^\infty dk_\perp M^2_{k_\perp,m}(E,E',E_t). \tag{A7}$$

The $t$-integral in $C(x, y)$ [Eq. (A5)] can be calculated exactly with the help of[35]

$$\int_{-\infty}^{\infty} dt Ai(x+t)Ai(\tilde{y}-\alpha t) = \frac{1}{(1+\alpha^3)^{\frac{1}{3}}}Ai\left(\frac{\tilde{y}+\alpha x}{(1+\alpha^3)^{1/3}}\right). \tag{A8}$$

Introducing Eq. (25) for the field-broadened DOS, the transition rate simplifies to

$$\tilde{D}_{\mathrm{t,c}}(E,E',E_\mathrm{t}) = 4\pi r_0^3 \varrho_\mathrm{t}(\tilde{E},E_\mathrm{t}) \frac{\theta_\mathrm{c}}{\theta_\mathrm{t}^3} \frac{(\mu^2 m_\mathrm{c})^{\frac{1}{3}}}{m_\mathrm{t}} \sqrt{E_\mathrm{t}\hbar\theta_\mathrm{t}} \int_0^\infty d\zeta |C'(\zeta)|^2 \quad \text{(A9)}$$

with

$$C'(\zeta) = \int_{-\infty}^{\infty} d\lambda \frac{\mathcal{P}}{\lambda} Ai\left(\lambda + \zeta + \frac{\tilde{E}}{\hbar\theta_\mathrm{t}}\right) \times Ai\left(a\lambda + \zeta/a^2 + \frac{\tilde{E}-E'}{\hbar\theta_\mu}\right). \quad \text{(A10)}$$

Here, $a = \theta_\mathrm{t}/\theta_\mu = (\mu/m_\mathrm{t})^{1/3}$ is the decisive parameter. There is no general solution to the principle value integral (A10) for an arbitrary $a$. An approximate treatment has been suggested by Bechstedt *et al.*[36] For demonstration, the abbreviations $p = \zeta + \tilde{E}/\hbar\theta_\mathrm{t}$ and $q = \zeta/a^2 + (\tilde{E}-E')/\hbar\theta_\mu$ are defined. Using the integral representation of the Airy function, $C'$ can be re-written as[36]

$$C' = \int_{-\infty}^{\infty} d\lambda \frac{\mathcal{P}}{\lambda} Ai(\lambda+p) Ai(a\lambda+q) = \frac{ia}{4\pi} \int_{-\infty}^{\infty} d\sigma \int_{-\infty}^{\infty} ds\, e^{-\frac{i}{3}a^3\sigma^3 - ia\sigma ap + \frac{i}{3}s^3 + isq} sgn(s-\sigma), \quad \text{(A11)}$$

where *sgn* denotes the signum function. Now, the case of *strong localization* is defined by the condition $a^3 \ll 1$ and the factor $\exp(-ia^3\sigma^3/3)$ is replaced by unity in the $\sigma$-integral. This results in

$$C' \to -\frac{\mathcal{P}}{p} Ai(q-ap) = -\frac{\mathcal{P}}{\zeta + \tilde{E}/\hbar\theta_\mathrm{t}} Ai\left(a\frac{m_\mathrm{t}}{m_\mathrm{c}}\zeta - \frac{E'}{\hbar\theta_\mu}\right). \quad \text{(A12)}$$

Inserting into (A9), setting $\zeta = 0$ in the denominator ($Ai^2$ is rapidly decaying at $E' = -E_\mathrm{g} + E_\mathrm{t}$), and introducing the joint DOS $\varrho_\mu(E')$ from Eq. (28) gives

$$\tilde{D}_{\mathrm{t,c}}(E,E',E_\mathrm{t}) = 16\pi^2 \frac{m_\mathrm{c}}{\mu} \frac{\mathcal{P}}{\tilde{E}^2} E_\mathrm{t}^2 r_0^6 \varrho_\mathrm{t}(\tilde{E},E_\mathrm{t}) \varrho_\mu(E'). \quad \text{(A13)}$$

The final step is to multiply the energetic transition rate for one localized state $\tilde{D}$ by the density of the energy levels $1/(2\pi r_0^3)$ which yields Eq. (27):

$$D_{\mathrm{t,c}}(E,E',E_\mathrm{t}) = 8\pi \frac{m_\mathrm{c}}{\mu} \frac{\mathcal{P}}{\tilde{E}^2} E_\mathrm{t}^2 r_0^3 \varrho_\mathrm{t}(\tilde{E},E_\mathrm{t}) \varrho_\mu(E'). \quad \text{(A14)}$$

The case of *very weak localization* is defined by the condition $a^3 \approx 1$.[37] One obtains for $C'$

$$C' \to a \int_{-\infty}^{\infty} d\lambda \frac{\mathcal{P}}{\lambda} Ai(ap+\lambda) Ai(q+\lambda) = -a\pi[\Theta(ap-q) Ai(ap) Bi(q) + \Theta(q-ap) Bi(ap) Ai(q)]. \quad \text{(A15)}$$

The first term in braces is zero because the argument of the $\Theta$-function is always negative. Thus,

$$\int_{-\infty}^{\infty} d\zeta |C'(\zeta)|^2 = a^2\pi^2 Bi^2\left(\frac{\tilde{E}}{\hbar\theta_\mu}\right) \int_{-\infty}^{\infty} d\zeta Ai^2\left(\frac{\zeta}{a^2} + \frac{E_\mathrm{g}}{\hbar\theta_\mu}\right) \approx a\pi^2 Bi^2\left(\frac{\tilde{E}}{\hbar\theta_\mu}\right) \mathcal{F}\left(\frac{E_\mathrm{g}}{\hbar\theta_\mu}\right) \quad \text{(A16)}$$

which finally results in Eq. (35)

$$D_{\mathrm{t,c}}(E,E',E_\mathrm{t}) = 8\pi \frac{Bi^2\left(\frac{\tilde{E}}{\hbar\theta_\mu}\right)}{(\hbar\theta_\mu)^2} \frac{m_\mathrm{t}}{\mu} \times E_\mathrm{t}^2 r_0^3 \varrho_\mathrm{t}(\tilde{E},E_\mathrm{t}) \varrho_\mu(-E_\mathrm{g}). \quad \text{(A17)}$$

## APPENDIX B: REPRESENTATION OF BAND-TO-BAND TUNNEL RATE BY CONVOLUTION OF IDEAL CB AND IDEAL VB DOSs

In this appendix it is shown that the ideal BTBT rate, which is proportional to the reduced DOS, can be represented as convolution of the ideal CB DOS and the ideal VB DOS.

Inserting the function $\mathcal{F}$ for both bands explicitly into Eq. (40), one obtains

$$G_\mathrm{BTB} = c \int_{-\infty}^{\infty} dE\, \varrho_\mathrm{v}(\tilde{E})\, \varrho_\mathrm{c}(E) = c\, \frac{(2m_\mathrm{v})^{3/2}(2m_\mathrm{c})^{3/2}}{(2\pi\hbar^3)^2} \sqrt{\hbar\theta_\mathrm{v}} \sqrt{\hbar\theta_\mathrm{c}}\, M, \quad \text{(B1)}$$

where

$$M = \int_{-\infty}^{\infty} dE \int_0^\infty dx \int_0^\infty dy\, Ai^2\left(x + \frac{\tilde{E}}{\hbar\theta_\mathrm{v}}\right) Ai^2\left(y - \frac{E}{\hbar\theta_\mathrm{c}}\right). \quad \text{(B2)}$$

With the help of[35]

$$\int_{-\infty}^{\infty} dt Ai^2(x+t) Ai^2(\tilde{y} - \alpha t) = \frac{1}{2\pi\sqrt{\alpha}} \int_0^\infty \frac{du}{\sqrt{u}} Ai^2\left(u + \frac{\tilde{y} + \alpha x}{(1+\alpha^3)^{1/3}}\right), \quad \text{(B3)}$$

the obvious identity

$$\int_0^\infty dx \int_0^\infty dy f(x+y) = \int_0^\infty dt\, t f(t), \quad \text{(B4)}$$

and

$$\int_0^\infty \frac{du}{\sqrt{u}} \int_0^\infty dv\, v Ai^2(u+v+x) = \frac{4}{3} \int_0^\infty dt\, t^{3/2} Ai^2(t+x), \quad \text{(B5)}$$

$M$ can be written as

$$M = \hbar\theta_r \frac{(m_c m_v)^{1/6}}{2\pi m_r^{1/3}} \frac{4}{3} \int_0^\infty dt\, t^{3/2} Ai^2\left(t + \frac{E_g}{\hbar\theta_r}\right). \tag{B6}$$

The integral in Eq. (B6) can be calculated exactly using a recursion relation[35] for $\int_0^\infty dt\, t^n Ai^2(t+x)$ which gives

$$\int_0^\infty dt\, t^{3/2} Ai^2(t+x) = \frac{3}{64}\left[4x^2 Ai_1(\kappa x) + \kappa^2 x Ai'(\kappa x) + \kappa Ai(\kappa x)\right], \quad \kappa = 2^{2/3}. \tag{B7}$$

With this, the parameter $c$ is determined but depends in a complicated way on $\mathcal{F}$, $Ai_1$, $Ai'$, and $Ai$. It takes a much simpler form if the asymptotic limits of these functions are used. This is admissible since $E_g \gg \hbar\theta_r$ holds in all practical cases. For the correct asymptotic limit of $M$, one has to expand $Ai_1$ and $Ai'$ up to third order and $Ai$ up to second order, all lower-order terms cancel each other. These asymptotic forms read (note a mistake for $Ai_1$ in Ref. 39 corrected by Nikishov and Ritus[38] and studied for BTBT in Ref. 40)

$$Ai_1(z) = \frac{z^{-3/4}}{2\sqrt{\pi}} e^{-\frac{2}{3}z^{3/2}}\left(1 - \frac{41}{48z^{3/2}} + \frac{9241}{4608z^3} - \cdots\right),$$
$$Ai'(z) = -\frac{z^{1/4}}{2\sqrt{\pi}} e^{-\frac{2}{3}z^{3/2}}\left(1 + \frac{7}{48z^{3/2}} - \frac{455}{4608z^3} - \cdots\right),$$
$$Ai(z) = \frac{z^{-1/4}}{2\sqrt{\pi}} e^{-\frac{2}{3}z^{3/2}}\left(1 - \frac{5}{48z^{3/2}} + \frac{385}{4608z^3} - \cdots\right).$$

Equation (B7) becomes in the asymptotic limit

$$\lim_{x\to\infty} \int_0^\infty dt\, t^{3/2} Ai^2(t+x) = \frac{3}{64\sqrt{2\pi}x^{7/4}} e^{-\frac{4}{3}x^{3/2}}. \tag{B8}$$

Now one easily finds Eq. (41) for the parameter $c$ using the asymptotic limit of $G_{BTB}$ in Eq. (B1).

## APPENDIX C: FIELD MODIFICATION OF CONTINUUM STATES OF THE NON-IDEAL DOS

Continuum states have energies $\tilde{E} < E_{edge}$ in the VB DOS, and the effective mass $m_t$ is assumed to be equal to the hole mass $m_v$. In this appendix, it is shown how the non-ideal VB DOS in this energy range changes under the influence of a constant electric field. The field broadening will be first demonstrated for the case of the ideal DOS

$$\varrho_v^{(F=0)}(\tilde{E}) = \frac{\sqrt{8m_v^3}}{4\pi^2\hbar^2}\sqrt{-\tilde{E}}\,\Theta(-\tilde{E}). \tag{C1}$$

As $\lim_{F\to 0}\varrho_v(\tilde{E}) = \varrho_v^{(F=0)}(\tilde{E})$, it follows:

$$\lim_{F\to 0}\frac{1}{\sqrt{\hbar\theta_v}}\int_{-\infty}^\infty d\epsilon\,\Theta(\epsilon)\,Ai^2\left(\frac{\epsilon+\tilde{E}}{\hbar\theta_v}\right) = \frac{1}{\pi}\int_{-\infty}^\infty d\epsilon\,\Theta(\epsilon)\sqrt{\epsilon}\,\delta(\epsilon+\tilde{E}). \tag{C2}$$

Therefore, one can enforce the field-broadening by the replacement

$$\Theta(\epsilon)\sqrt{\epsilon}\,\delta(\epsilon+\tilde{E}) \to \frac{\Theta(\epsilon)\pi}{\sqrt{\hbar\theta_v}} Ai^2\left(\frac{\epsilon+\tilde{E}}{\hbar\theta_v}\right). \tag{C3}$$

This is now done for the non-ideal DOS in the energy range $\tilde{E} < E_{edge}$. It is given by

$$\varrho_{v,t}^{(F=0)}(\tilde{E}) = \frac{(2m_v)^{3/2}\sqrt{\eta}}{2\pi^2\hbar^3} Y\left(-\frac{\tilde{E}}{\eta}\right)\Theta(E_{edge}-\tilde{E}) = \frac{(2m_v)^{3/2}\sqrt{\eta}}{2\pi^2\hbar^3}\int_{-\infty}^\infty d\epsilon\,\delta(\epsilon+\tilde{E})\,Y\left(\frac{\epsilon}{\eta}\right)\Theta(E_{edge}+\epsilon). \tag{C4}$$

After shifting the $\epsilon$-integration by $-E_{edge}$ one can immediately apply the replacement (C3) which yields Eq. (42).

## APPENDIX D: TRANSITION RATE FROM NON-IDEAL CONTINUUM STATES

This appendix presents the calculation of the rate of transitions from continuum states of the non-ideal VB DOS into CB states.

Inserting Eq. (42) instead of $\varrho_v(\tilde{E})$ into the convolution integral Eq. (40), one can calculate the $E$-integral over $Ai^2Ai^2$ as done in Appendix B. Then

$$G_{tc}^{cond} = c\,\frac{(m_v m_c)^{3/2}\hbar\theta_r}{\pi^3\hbar^6}\sqrt{\eta}\int_0^\infty \frac{d\epsilon}{\sqrt{\epsilon}} Y\left(\frac{\epsilon - E_{edge}}{\eta}\right) \times \int_0^\infty dv \int_0^\infty \frac{du}{\sqrt{u}}\, Ai^2\left(u+v+\frac{\epsilon+E_g'}{\hbar\theta_r}\right). \tag{D1}$$

Using the identity

$$\int_0^\infty \frac{du}{\sqrt{u}}\int_0^\infty dv\, Ai^2(u+v+x) = 2\int_0^\infty dt\,\sqrt{t}Ai^2(t+x), \tag{D2}$$

and inserting $c$ results into Eq. (43)

$$G_{tc}^{cond} = \frac{(eF)^3}{2\hbar E_g}\left(\frac{E_g}{\hbar\theta_r}\right)^{\frac{3}{4}}\frac{\sqrt{\pi}\sqrt{\eta}}{(\hbar\theta_r)^2}\int_0^\infty \frac{d\epsilon}{\sqrt{\epsilon}} Y\left(\frac{\epsilon - E_{edge}}{\eta}\right) \times \int_0^\infty dt\,\sqrt{t}Ai^2\left(t+\frac{\epsilon+E_g'}{\hbar\theta_r}\right). \tag{D3}$$

The last integral can be exactly expressed by Airy functions using recursion relations[35]

$$\int_0^\infty dt\,\sqrt{t}Ai^2(t+x) = -\frac{\kappa^2}{16}Ai'(\kappa x) - \frac{1}{4}x\,Ai_1(\kappa x). \tag{D4}$$

To make it proportional to $\mathcal{F}$, the ansatz

$$\int_0^\infty dt\,\sqrt{t}Ai^2(t+x) = b(x)\,\mathcal{F}(x) \tag{D5}$$

is invoked. In the asymptotic limit ($E_g \gg \hbar\theta_r$), Eq. (D4) yields

$$\lim_{x\to\infty}\int_0^\infty dt\,\sqrt{t}Ai^2(t+x) = \frac{x^{-5/4}}{16\sqrt{2\pi}}\,e^{-\frac{4}{3}x^{3/2}}, \tag{D6}$$

and, therefore, with Eq. (D5)

$$\lim_{x\to\infty} b(x)\,\mathcal{F}(x) = \frac{1}{8\pi x}\,e^{-\frac{4}{3}x^{3/2}}\lim_{x\to\infty} b(x), \tag{D7}$$

and by comparison of Eqs. (D6) and (D7) one finds $b(x)$ for large $x$:

$$b(x) = \frac{\sqrt{2\pi}}{4x^{1/4}}. \tag{D8}$$

Now, Eq. (D3) can be written as

$$G_{tc}^{cond} = \frac{(eF)^3}{8\hbar E_g}\left(\frac{E_g}{\hbar\theta_r}\right)^{\frac{3}{4}}\frac{\sqrt{2}\pi\sqrt{\eta}}{(\hbar\theta_r)^2}\int_0^\infty\frac{d\epsilon}{\sqrt{\epsilon}}Y\left(\frac{\epsilon-E_{edge}}{\eta}\right) \times\left(\frac{\hbar\theta_r}{\epsilon+E_g'}\right)^{\frac{1}{4}}\mathcal{F}\left(\frac{\epsilon+E_g'}{\hbar\theta_r}\right). \tag{D9}$$

The slowly varying factor $(\ldots)^{1/4}$ in the last line can be taken in front of the integral at $\epsilon = E_{edge}$. Then, replacing $\mathcal{F}$ by the rate of direct BTBT

$$\mathcal{F}\left(\frac{\epsilon+E_g'}{\hbar\theta_r}\right) = \frac{8\hbar E_g\hbar\theta_r}{(eF)^3}G_{BTB}\left(\frac{\epsilon+E_g'}{\hbar\theta_r}\right) \tag{D10}$$

one obtains Eqs. (44) and (45). On the other hand, using the asymptotic limit for $\mathcal{F}$, Eq. (D9) turns into

$$G_{tc}^{cond} = \frac{(eF)^3}{64\hbar E_g}\left(\frac{\hbar\theta_r}{E_g}\right)^{\frac{1}{2}}\frac{\sqrt{2}\sqrt{\eta}}{(\hbar\theta_r)^2}\int_0^\infty\frac{d\epsilon}{\sqrt{\epsilon}}Y\left(\frac{\epsilon-E_{edge}}{\eta}\right) \times\exp\left[-\frac{4}{3}\left(\frac{\epsilon+E_g'}{\hbar\theta_r}\right)^{3/2}\right]. \tag{D11}$$

Expanding the exponent for small $\epsilon$ and performing a partial integration wrt the factor $1/\sqrt{\epsilon}$, this can be written as

$$G_{tc}^{cond} = -\frac{(eF)^3}{32\hbar E_g}\left(\frac{\hbar\theta_r}{E_g}\right)^{\frac{1}{2}}\frac{\sqrt{2}\sqrt{\eta}}{(\hbar\theta_r)^2}\exp\left[-\frac{4}{3}\left(\frac{E_g'}{\hbar\theta_r}\right)^{3/2}\right] \times\int_0^\infty d\epsilon\sqrt{\epsilon}\left[\frac{1}{\eta}Y'\left(\frac{\epsilon-E_{edge}}{\eta}\right)-\frac{2\sqrt{E_g'}}{(\hbar\theta_r)^{3/2}}Y\left(\frac{\epsilon-E_{edge}}{\eta}\right)\right]\exp\left(-\frac{2\sqrt{E_g'}}{(\hbar\theta_r)^{3/2}}\epsilon\right). \tag{D12}$$

The second term in angular braces dominates due to its proportionality to $(E_g'/\hbar\theta_r)^{1/2}$. Retaining only this term, taking $Y[(\epsilon - E_{edge})/\eta]$ out of the integral at $\epsilon^* = (\hbar\theta_r)^{3/2}/(4\sqrt{E_g'})$, where the rest of the integrand becomes maximum, gives

$$G_{tc}^{cond} = \frac{(eF)^3}{64\hbar E_g^2}\frac{\sqrt{\pi}\sqrt{\eta}E_g^{1/4}}{(\hbar\theta_r)^{3/4}} \times Y\left[\frac{(\hbar\theta_r)^{3/2}}{4\eta E_g^{1/2}}-\frac{E_{edge}}{\eta}\right]e^{-\frac{4}{3}\left(\frac{E_g'}{\hbar\theta_r}\right)^{\frac{3}{2}}}. \tag{D13}$$

To obtain the limit for the ideal DOS ($\eta \to 0$), one has to multiply by $\pi^{3/2}/2$ which corrects for the above approximations. This finally yields Eq. (45).


[1] W. C. Vinogradov, Fiz. Tverd. Tela **13**, 3266 (1971); Sov. Phys. - Solid State **13**(11), 2745 (1972).
[2] S. Banerjee, W. Richardson, J. Coleman, and A. Chatterjee, IEEE Electron Dev. Lett. **8**, 347 (1987).
[3] A. M. Ionescu and H. Riel, Nature **479**, 329 (2011).
[4] M. A. Kayer and R. Lake, J. Appl. Phys. **110**, 074508 (2011).
[5] A. Schenk, S. Sant, K. Moselund, and H. Riel, in *Proceedings of the ULIS-EUROSOI* (2016), pp. 9–12.
[6] E. Memisevic, M. Hellenbrand, E. Lind, A. R. Persson, S. Sant, A. Schenk, J. Svensson, R. Wallenberg, and L.-E. Wernersson, Nano Lett. **17**(7), 4373 (2017).
[7] M. Takeshima, Phys. Rev. B **12**(2), 575 (1975).
[8] J. Teherani, S. Agarwal, W. Chern, P. Solomon, E. Yablonovitch, and D. Antoniadis, J. Appl. Phys. **120**, 084507 (2016).
[9] A. L. Efros, Sov. Phys.-Usp. **16**, 789 (1974).
[10] P. A. Wolff, Phys. Rev. **126**, 405 (1962).
[11] P. Chakraborty and K. Ghatak, Phys. Lett. A **288**, 335 (2001).
[12] S. Sylvia, M. Khayer, K. Alam, and R. Lake, IEEE Trans. Electron Devices **59**, 2996 (2012).
[13] S. Sylvia, K. Habib, M. Khayer, K. Alam, M. Neupane, and R. Lake, IEEE Trans. Electron Devices **61**, 2208 (2014).
[14] E. O. Kane, Solid-State Electron. **28**, 3 (1985).
[15] E. O. Kane, Phys. Rev. **131**, 79 (1963).
[16] G. Lucovsky, Solid State Commun. **3**, 299 (1965).
[17] A. Schenk, R. Enderlein, and D. Suisky, Phys. Status Solidi B **131**, 729 (1985).
[18] S. N. Mott, J. Phys. C: Solid State Phys. **20**, 3075 (1987).
[19] A. Schenk, M. Stahl, and H.-J. Wünsche, Phys. Status Solidi B **154**, 815 (1989).
[20] B. I. Halperin and M. Lax, Phys. Rev. **148**, 722 (1966).
[21] B. I. Shkiovskii and A. L. Efros, Sov. Phys.- Semicond. **4**, 249 (1970).
[22] J. Katahara and H. Hillhouse, J. Appl. Phys. **116**, 173504 (2014).
[23] J. R. Dixon and J. M. Ellis, Phys. Rev. **123**, 1560 (1961).
[24] V. Malyutenko and V. Chernyakhovsky, Semicond. Sci. Techol. **9**, 1047 (1994).
[25] S. W. Kurnick and J. M. Powell, Phys. Rev. **116**, 597 (1959).
[26] J. I. Pankove, Phys. Rev. **140**, A2059 (1965).
[27] D. Redfield, J. Non-Cryst. Solids **8–10**, 602 (1972).
[28] Synopsys Inc., Sentaurus Device User Guide, Version- 2015.06, Mountain View, California, 2015.
[29] E. O. Kane, J. Phys. Chem. Solids **12**, 181 (1959).
[30] E. O. Kane, J. Appl. Phys. **32**, 83 (1961).
[31] H. Carrillo-Nunez, A. Ziegler, M. Luisier, and A. Schenk, J. Appl. Phys. **117**, 234501 (2015).
[32] L. V. Keldysh, Sov. Phys. JETP **6**(33), 763 (1958).
[33] see http://www.ioffe.ru/SVA/NSM/Semicond/InAs for the band structure quantities such as the effective masses and the band gap listed along with the references.
[34] M. Luisier and G. Klimeck, J. Appl. Phys. **107**, 084507 (2010).
[35] D. E. Aspnes, Phys. Rev. **147**, 554 (1966).
[36] F. Bechstedt, R. Enderlein, and K. Peuker, Phys. Status Solidi B **68**, 43 (1975).

[37]R. Enderlein, J. Fiddicke, F. Bechstedt, K. Peuker, and R. S. Bauer, in Proceedings of the Luminescence Conference, Berlin (West), July 20–24 1981.
[38]A. I. Nikishov and V. I. Ritus, "Asymptotic representations for some functions and integrals connected with the Airy function," e-print arXiv:math-ph/0501062v1.
[39]M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover Publications, New York, 1970), p. 449.
[40]A. Heigl, A. Schenk, and G. Wachutka, in *Proceedings of SISPAD (*2009*)*, pp. 267–268.
[41]S. Sant and A. Schenk, J. Electron. Devices Soc. **3**(3), 164 (2015).