Efficient Modeling of Charge Trapping at Cryogenic Temperatures—Part I: Theory

Jakob Michl, Alexander Grill, Dominic Waldhoer, Wolfgang Goes, Ben Kaczer, Dimitri Linten, *Senior Member, IEEE*, Bertrand Parvais, Bogdan Govoreanu, *Senior Member, IEEE*, Iuliana Radu, Michael Waltl, *Member, IEEE*, and Tibor Grasser, *Fellow, IEEE*

***Abstract*— Charge trapping is arguably the most important detrimental mechanism distorting the ideal characteristics of MOS transistors, and nonradiative multiphonon (NMP) models have been demonstrated to provide a very accurate description. For the calculation of the NMP rates at room temperature or above, simple semiclassical approximations have been successfully used to describe this intricate mechanism. However, for the computation of charge transition rates at cryogenic temperatures, it is necessary to use the full quantum mechanical description based on Fermi's golden rule. Since this is computationally expensive and often not feasible, we discuss an efficient method based on the Wentzel–Kramers–Brillouin (WKB) approximation in combination with the saddle point method and benchmark this approximation against the full model. We show that the approximation delivers excellent results and can, hence, be used to model charge trapping behavior at cryogenic temperatures.**

***Index Terms*— Advanced CMOS, bias temperature instability (BTI), cryo-CMOS, cryogenic, physical modeling.**

I. Introduction

ADVANCED CMOS technologies operated at cryogenic temperatures provide the backbone of the space industry, astronomy, and high-precision metrology [1]. They are also required to enable large-scale quantum computing, which has to date an operation range of 4 K down to a few mK [2]–[4]. Currently, the scalability of the quantum-classical interface and the corresponding readout and control units are the biggest barrier to build quantum computers with a large number of qubits, and it is likely that this challenge remains even when it is possible to process physical qubits with sufficient fidelity to construct logical qubits with low error rates [5].

Recent publications on cryo-CMOS technologies mainly focus on the behavior of important time-zero device characteristics, such as subthreshold swing, ON-state current, and threshold voltage [6]–[8]. These time-zero characteristics show that, at cryogenic temperatures, the electrical performance of MOSFETs typically improves (higher ON-current and transconductance, steeper subthreshold slope, lower leak current, and so on). Consequently, MOSFETs can be operated using lower voltages compared to room temperature applications, which partially mitigates the risk of reliability issues. However, recent measurements show that, still, even at cryogenic temperatures, reliability issues, such as hot carrier degradation (HCD) [9], bias temperature instability (BTI) [10], and random telegraph noise (RTN) [11], [12], play a role and have to be carefully addressed for cryo-CMOS circuits.

The nonradiative multiphonon (NMP) theory has been used successfully to explain charge trapping phenomena for a wide

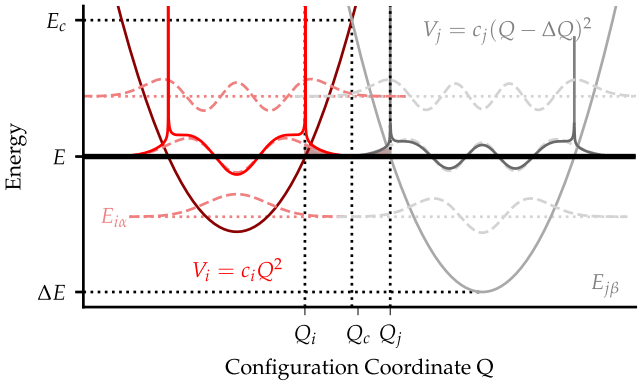

Fig. 1. For the computation of the transition rates for SEPC between two states represented by two PECs V_i and V_j, it is necessary to find the vibrational wave functions occurring in Fermi's golden rule (dashed red/gray lines). These can be approximated using the WKB approximation (solid red/gray lines) developed at the classical turning points Q_i and Q_j.

Manuscript received July 15, 2021; revised September 21, 2021; accepted September 22, 2021. Date of publication October 27, 2021; date of current version December 1, 2021. This work was supported in part by imec's Industrial Affiliation Program on Quantum Computing and Cryoelectronics, in part by the Austrian Research Promotion Agency FFG (Take Off Programme) under Project 861022 and Project 867414, in part by the European Union's Horizon 2020 Research and Innovation Program under Grant Agreement 871813, within the framework of the project Modeling Unconventional Nanoscaled DeviceFABrication (MUNDFAB), in part by the Austrian Federal Ministry for Digital and Economic Affairs, in part by the National Foundation for Research, Technology and Development, and in part by the Christian Doppler Research Association. The review of this article was arranged by Editor E. A. Gutiérrez-D. *(Corresponding author: Jakob Michl.)*

Jakob Michl, Dominic Waldhoer, and Tibor Grasser are with the Institute for Microelectronics, TU Wien, 1040 Vienna, Austria (e-mail: michl@iue.tuwien.ac.at).

Alexander Grill, Ben Kaczer, Dimitri Linten, Bogdan Govoreanu, and Iuliana Radu are with imec, 3001 Leuven, Belgium.

Wolfgang Goes is with Silvaco Europe Ltd., Cambridge PE27 5JL, U.K.

Bertrand Parvais is with imec, 3001 Leuven, Belgium, and also with the Electronics and Informatics (ETRO) Department, Vrije Universiteit Brussel, 1050 Brussels, Belgium.

Michael Waltl is with the Christian Doppler Laboratory for Single Defect Spectroscopy, Institute for Microelectronics, TU Wien, 1040 Vienna, Austria.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TED.2021.3116931.

Digital Object Identifier 10.1109/TED.2021.3116931

range of technologies and temperatures [13]. Hereby, charge trapping occurs either in intrinsic defects, e.g., strained bonds and oxygen vacancies in SiO_2 [14] and HfO_2 [15], or extrinsic defects, e.g., hydrogen-related defects [16], [17]. For the computation of the charge transition rates, approximations are made, which typically hold true in the limit of high temperatures, but are not valid at cryogenic temperatures anymore. In a quantum mechanical picture, the nuclei wave functions of the initial and final states overlap, which is interpreted as nuclear tunneling process from one atomistic configuration to another [18]. Thus, oxide defects can still be active in a low-temperature regime, and the explicit calculation of vibrational wave function overlaps becomes necessary [19]. Since this is computationally expensive, it is necessary to find an approximation feasible for efficient TCAD modeling. In this work, we present a computationally efficient model based on the Wentzel–Kramers–Brillouin (WKB) approximation combined with the saddle point method (SPM), which can be used for the simulation of charge trapping down to cryogenic temperatures.

II. Nonradiative Multiphonon Model

Simple and efficient two-state NMP models have been regularly used to describe the dynamics of charging and discharging of oxide defects at room temperature and above [13], [20]. Within the Born–Oppenheimer approximation, both the neutral and charged defect states can be described by separate potential energy curves (PECs), which are typically approximated as harmonic oscillators. Fig. 1 shows the charge state i described by the potential energy V_i and the neutral state j described by the potential energy V_j. In the limit of high temperatures, a transition between the charge states is possible by overcoming the classical barrier, which is determined by the intersection point (IP) of the PECs. Within this approximation, the transition rates take the simple analytical form [13]

$$k_{ij} = n v_{\mathrm{th}} \vartheta_{\mathrm{WKB}} \sigma \exp\left(-\frac{\epsilon_{ij}}{k_{\mathrm{B}}T}\right) \tag{1}$$

with the channel carrier concentration n, the thermal velocity v_{th}, the tunneling factor ϑ_{WKB}, which can be computed using WKB approximation [21], the capture cross section σ [22], and the barrier ϵ_{ij}. The latter can be computed using the Huang–Rhys parameters [23]: relaxation energy E_{R}, trap level E_{T}, and the ratio R of the curvatures of the two PECs. The trap level E_{T} is used to compute the energetic offset ΔE to the band edges at a given gate bias V_{G}.

In the quantum mechanical picture, the transition rate from charge state i to state j has to be computed using Fermi's golden rule

$$k_{ij}(T) = \frac{2\pi}{\hbar}\frac{|\theta_{ij}|^2}{Z}\sum_{\alpha\beta}|I_{i\alpha,j\beta}|^2\delta(E_{i\alpha}-E_{j\beta})e^{-E_{i\alpha}/k_{\mathrm{B}}T} \tag{2}$$

where θ_{ij} is the electronic matrix element, and $E_{i\alpha}$ and $E_{j\beta}$ are the vibrational eigenenergies of charge state i and j, while α and β are the indices of the eigenenergies. $\eta_{i\alpha}$ and $\eta_{j\beta}$ are the corresponding vibrational wave functions, and $I_{i\alpha,j\beta}$ is the overlap integral of these vibrational wave functions [19]. Z is the canonical partition function defined by

$$Z = \sum_{\gamma} e^{-E_{i\gamma}/k_{\mathrm{B}}T} \tag{3}$$

where γ runs over all accessible states.

The idea of (2) is to sum over all combinations of eigenstates α and β and compute the overlap of the wave functions in $I_{i\alpha,j\beta}$. These overlaps are then multiplied by the probability $e^{-E_{i\alpha}}/Z$ of the initial state being occupied. For this, it is assumed that the defect can exchange energy with the channel, and they are in thermal equilibrium.

Compared to the simple semiclassical expression (1), the full quantum mechanical summation (2) is computationally expensive because, first, it is necessary to compute the overlap integrals with the given eigenfunctions and eigenenergies, and afterward, the summation over all α and β needs to be performed. Even if optimized schemes for the computation of the overlap integrals are used, as, e.g., proposed by Schmidt [24], the full quantum mechanical summation is still computationally expensive compared to the WKB approximation.

A. Derivation of the WKB Approximation

An approximation for the quantum mechanical rate equation (2) was proposed by Holstein [25] and further discussed by Markvart [26]. Since these papers were restricted to weak-electron–phonon coupling, while a lot of experimental data require strong-electron–phonon coupling, a more general formulation is developed in the following. For that, we look at a single transition $\alpha \rightarrow \beta$ from Fermi's golden rule (2) and introduce the line shape function

$$\xi_{i\alpha,j\beta} = e^{-\beta\Omega_{i\alpha}}|I_{i\alpha,j\beta}|^2\delta(E_{i\alpha}-E_{j\beta}) \tag{4}$$

where the vibrational overlap integral $I_{i\alpha,j\beta}$ is given by

$$I_{i\alpha,j\beta} = \left|\int_{-\infty}^{\infty}\eta_{i\alpha}(Q)\eta_{j\beta}(Q)\mathrm{d}Q\right|^2. \tag{5}$$

In order to solve (5), the wave function $\eta_{i\alpha}(Q)$ will be approximated using the WKB approximation [21]. The WKB approximation is a semiclassical series expansion of a corresponding action with respect to $\hbar$ and, thus, needs to be carried out at a certain point, typically done at the classical turning point Q_i, as compared in Fig. 1. The expansion inside the classically allowed regions gives an oscillating function, which cancels when integrated and, thus, can be neglected [27]. Therefore, the WKB wave function reduces to an exponentially decaying function in the classically prohibited region

$$\eta_{i\alpha}(Q) = \frac{(-1)^{\alpha}C_i}{\sqrt{|k_{i\alpha}(Q,E_{i\alpha})|}}e^{\frac{1}{\hbar}\int_{Q_i}^{Q}k_{i\alpha}(Q',E_{i\alpha})\mathrm{d}Q'} \tag{6}$$

with $k_{i\alpha}(Q,E_{i\alpha}) = (2m(U_i(Q)-E_{i\alpha}))^{1/2}$ being the classical momentum of the particle. The constant C_i can be computed by normalizing $\eta_{i\alpha}(Q)$ giving $C_i = (m_i\omega_i/8\pi^2)^{(1/4)}$, with ω_i being the oscillator frequency [27].

Fig. 1 shows that, in the case of strong-electron–phonon coupling, the classically prohibited parts of the wave functions between Q_i and Q_j dominate. An overview of all occurring

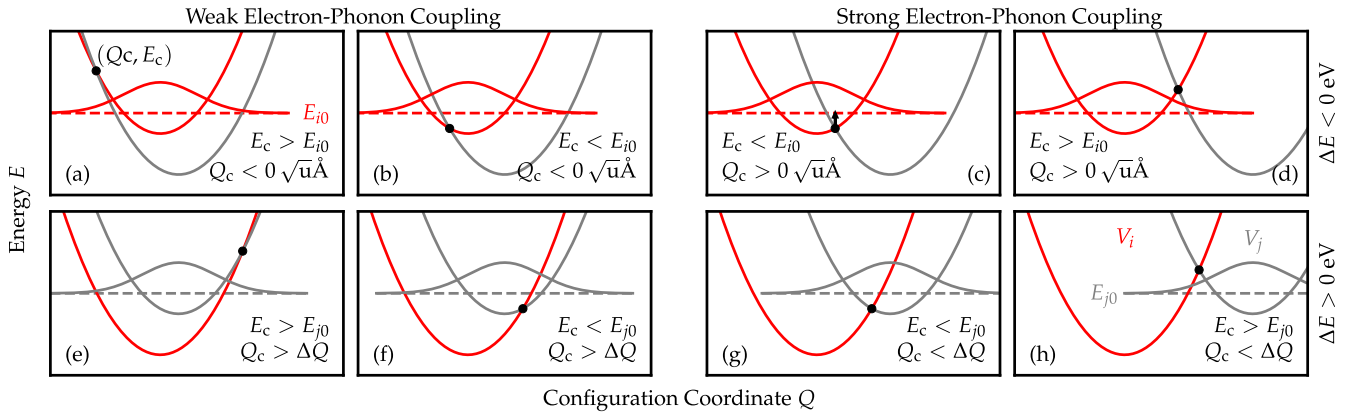

Fig. 2. There are eight different relative positions of the PECs that need to be examined separately. (a)–(d) Defining the minimum of PEC i (red) to be at $Q = 0\ \sqrt{\text{u}}\text{Å}$, there are four cases with $\Delta E < 0$ eV. (a) and (b) Weak electron–phonon coupling (WEPC), with the difference that, in (a), the IP is above the ground state E_{i0} and, in (b) it is below. (c) and (d) SEPC, with IP (c) below and (d) above ground state of V_i. The lower row describes the relative positions with $\Delta E > 0$ eV. (e) and (f) SEPC with IP above and below the ground state of V_j, respectively. (g) and (h) WEPC case.

relative positions of the parabolas, where an IP occurs, can be seen in Fig. 2.

The evaluation of the overlap integrals (5) is performed using the SPM [28]. For that, we express the overlap integral using the WKB functions

$$I_{i\alpha,j\beta} = \left| \frac{C_i C_j}{\sqrt{|k_{i\alpha}(Q, E_{i\alpha})|}\sqrt{|k_{j\beta}(Q, E_{j\beta})|}} e^{\frac{\varphi_{i\alpha j\beta}(Q, E_{i\alpha}, E_{j\beta})}{2}} \right|^2 \quad (7)$$

with the phase $\varphi_{i\alpha j\beta}(Q, E_{i\alpha}, E_{j\beta})$ given by

$$\varphi_{i\alpha j\beta}(Q, E_{i\alpha}, E_{j\beta}) = \frac{2}{\hbar}\int_{Q_i}^{Q} k_{i\alpha}(Q', E_{i\alpha})\mathrm{d}Q' - \frac{2}{\hbar}\int_{Q_j}^{Q} k_{j\beta}(Q', E_{j\beta})\mathrm{d}Q'. \quad (8)$$

The SPM exploits the fact that the main contribution to the integral stems from the region near the point of stationary phase, which is given by Q_c, as can be easily verified by solving for $\mathrm{d}\varphi_{i\alpha j\beta}(Q)/\mathrm{d}Q = 0$. Neglecting the Q dependence of the nonexponential prefactor, this results in

$$I_{i\alpha,j\beta}(E_{i\alpha}, E_{j\beta}) = \left| \frac{C_i C_j}{\sqrt{|k_{i\alpha}(Q_c, E_{i\alpha})|}\sqrt{|k_{j\beta}(Q_c, E_{j\beta})|}} \times e^{\frac{\varphi_{i\alpha j\beta}(Q_c, E_{i\alpha}, E_{j\beta})}{2}} \sqrt{\frac{\pi}{\varphi''_{i\alpha j\beta}(Q_c, E_{i\alpha}, E_{j\beta})}} \right|^2. \quad (9)$$

After the simplification of the overlap matrix via (9), the quantized line shape function (4) can be converted to

$$\xi_{i,j}(E, E') = e^{-\beta E_i}|I_{i,j}(E, E')|^2\delta(E - E') \quad (10)$$

using $E_{i\alpha} \to E$ and $E_{j\beta} \to E'$. Next the summation over the eigenenergies in (2) is converted to a continuous form using

$$\sum_{\alpha=0}^{\infty} f(E_\alpha) \to \int_{E_0}^{\infty} f(E)\frac{\mathrm{d}n\alpha}{\mathrm{d}E}\mathrm{d}E \quad (11)$$

where $\mathrm{d}\alpha/\mathrm{d}E$ is the density of states of the quantum mechanical harmonic oscillator

$$\frac{\mathrm{d}\alpha}{\mathrm{d}E} = \frac{\mathrm{d}}{\mathrm{d}E}\left(\frac{E}{\hbar\omega} - \frac{1}{2}\right) = \frac{1}{\hbar\omega}. \quad (12)$$

With the approximated line shape function (10), the rate equation (2) can be expressed as

$$k_{ij}(T) = \frac{2\pi}{\hbar}\frac{|\theta_{ij}|^2}{Z}\iint e^{-E/k_\mathrm{B}T}|I_{i,j}(E, E')|^2 \delta(E - E')\frac{1}{\hbar\omega_i}\frac{1}{\hbar\omega_j}\mathrm{d}E\mathrm{d}E'. \quad (13)$$

Here, Z now denotes the partition function of the quantum harmonic oscillator $Z = \exp(\hbar\omega_i/2k_\mathrm{B}T)/(1-\exp(\hbar\omega_i/k_\mathrm{B}T))$. Evaluating the Dirac delta distribution and using (9), (13) simplifies to

$$k_{ij}(T) = \int C_2(E)e^{-E/k_\mathrm{B}T + \varphi(E, Q_c)}\mathrm{d}E \quad (14)$$

with $\varphi(E, Q_c)$ being the continuous form of $\varphi_{\alpha\beta}(Q_c)$. $C_2(E)$ contains all nonexponential terms depending on E

$$C_2(E) = \frac{2\pi}{\hbar}\frac{|\theta_{ij}|^2}{Z}\frac{1}{\hbar\omega_i}\frac{1}{\hbar\omega_j}\left|\frac{C_i C_j}{\sqrt{2m(E_c - E)}}\right|^2\frac{2\pi}{\varphi''(Q_c)}. \quad (15)$$

The integral (14) can be evaluated using the SPM after finding the extremum E^*

$$\left.\frac{\mathrm{d}\varphi(E)}{\mathrm{d}E}\right|_{E=E^*} = \frac{1}{k_\mathrm{B}T} \quad (16)$$

leading to the result

$$k_{ij}(T) = C_2(E^*)e^{-E^*/k_\mathrm{B}T + \varphi(E^*)}\sqrt{\frac{2\pi}{\varphi''(E^*)}} \quad (17)$$

which, after insertion of all terms, yields the final result

$$k_{ij}(T) = \frac{2\pi}{\hbar}\frac{|\theta_{ij}|^2}{Z}\frac{1}{\hbar\omega_i}\frac{1}{\hbar\omega_j}\left|\frac{C_i C_j}{\sqrt{2m(E_c - E^*)}}\right|^2 \times e^{-E^*/k_\mathrm{B}T + \varphi(E^*, Q_c)}\sqrt{\frac{2\pi}{\varphi''(E^*)}}\frac{2\pi}{\mathrm{d}^2\varphi/\mathrm{d}Q^2(Q_c)}. \quad (18)$$

TABLE I
SIGNS IN ANALYTIC EXPRESSION OF THE OVERLAP INTEGRAL

	r_1	r_2	s_1	s_2
$Q_c < 0$	1	-1	-1	-1
$0 < Q_c < \Delta Q$	1	1	1	-1
$\Delta Q < Q_c$	-1	1	1	1

For the calculation of the dynamic response of MOSFETs to changing bias conditions, (18) will have to be repeatedly evaluated for every single defect at every single time step. It is possible to find analytic expressions for the integrals occurring in (8)

$$\begin{aligned}\varphi_{i\alpha j\beta}(Q_c, E_\alpha, E_\beta) &= \frac{-\sqrt{2}}{\hbar}\Bigg(r_1\Bigg(\frac{\sqrt{E_c}\sqrt{E_c - E_\alpha}}{\sqrt{c_i}} \\ &\quad - \frac{E_\alpha}{\sqrt{c_i}} \ln \frac{\sqrt{E_c - E_\alpha} + r_2\sqrt{E_c}}{r_2\sqrt{E_\alpha}}\Bigg) \\ &\quad + s_1\Bigg(\frac{\sqrt{E_c - \Delta E}\sqrt{E_c - E_\beta}}{\sqrt{c_j}} \\ &\quad - \frac{E_\beta - \Delta E}{\sqrt{c_j}} \ln \frac{\sqrt{E_c - E_\beta} + s_2\sqrt{E_c - \Delta E}}{s_2\sqrt{E_\beta - \Delta E}}\Bigg)\Bigg)\end{aligned} \tag{19}$$

which then allows to find analytic expressions for the first and second derivatives in direction Q. Here, the signs r_1, r_2, s_1, and s_2 differ depending on weak or strong electron–phonon coupling (SEPC) and are listed in Table I. Furthermore, the second derivative in direction of E can be found, which is needed in (16). Thus, the computational bottleneck is finding the minimum of (16), which needs to be done numerically, for example, by applying Newton's method.

B. Defining the Minimal Barrier

At cryogenic temperatures, the transition rate is dominated by the ground state wave function $\eta_{i/j0}$ and becomes temperature independent. For the simulation of this effect using the WKB approximation, it is necessary to define a minimal barrier. This can be seen by taking the limit $T \to 0$ K of the partition function Z. At this limit, the only occupied state is the ground state, and thus, the sum in Z reduces to $\exp(-E_{i0}/k_BT)$. In order for the transition rate to stay nonzero, the partition function needs to cancel out with the term $\exp(-E^*/k_BT)$ of the WKB approximation. To guarantee this, E^* must lie above the higher of the two ground states; otherwise, Z would diverge in the limit. For $E_c > E_{i/j0}$, as shown in Fig. 2(a), (d), (e), and (h), E^* can be directly computed using (16). For the case $E_c < E_{i/j0}$, E^* is fixed to the ground state with the higher energy. For the applied SPM for the integration over $\mathrm{d}Q$, we still use the extremum Q_c. Only for the second use of the SPM, $E_{i/j0}$ is used instead of $E^\star$, as shown in (18).

C. Discontinuity at the Turning Point

To guarantee a broad region in which the WKB approximation is valid, it is necessary to consider another special case. The WKB approximation is developed at the turning points Q_i and Q_j in Fig. 1, but, at these turning points, the wave functions diverge. In most cases, the IP E_c is far away from the turning point; however, in the special case $E^* = E_c$, this would lead to infinite transition rates, as can be seen in (18). This occurs when $E_c = E_{i0}$: in this case, $E^* < E_{i0}$ would hold, and due to section II-B, $E^* = E_{i0}$ is fixed, which would lead to the divergence.

To avoid this, the diverging terms of the form $1/(E_c - E^*)^{1/2}$, where $E_c \to E_{i0}$ and $E^* = E_{i0}$, can be substituted by

$$\begin{aligned}\frac{1}{\sqrt{|E_c - E^*|}} &= \frac{1}{\sqrt{|E_c - E_{i0}|}} \\ &\approx \frac{1}{\sqrt[4]{|E_c - E_{i0} + \epsilon|}\sqrt[4]{|E_c - E_{i0} - \epsilon|}}.\end{aligned} \tag{20}$$

The value of ϵ should be chosen in a way that $E_{i0} + \epsilon < E_{i1}$. In this work, we chose $\epsilon = \hbar\omega_i/2$.

III. COMPARISON FULL QM MODEL VERSUS WKB/SPM MODEL

In the following, we compare our WKB approximation derived above to the full quantum mechanical model presented in [19]. The sensitivity to the NMP parameters is investigated by spanning a grid of E_R, R, ΔE, and ΔQ values and observing the resulting error distributions. We show that the approximation error is negligible throughout the entire relevant parameter space.

A. Lifetime Broadening

Within the full QM model, it is necessary to introduce a lifetime broadening of the Dirac delta distribution $\delta(E_{i\alpha} - E_{j\beta})$ in (2) [29]. Without this broadening, there would be only a nonzero result for a perfect energetic alignment $E_{i\alpha} = E_{j\beta}$, which, in general, is not the case. Therefore, the delta distribution is replaced by a Gaussian distribution

$$\delta(E_{i\alpha} - E_{j\beta}) \approx \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(E_{i\alpha} - E_{j\beta})^2}{2\sigma^2}}. \tag{21}$$

It is obvious that the choice of the width σ of the distribution is crucial. Since the broadening is a thermal effect, a temperature-dependent σ of the form $\sigma = k_BT$ would appear reasonable. However, even at low temperatures, the excited states show a finite lifetime due to phonon emission and are, thus, affected by lifetime broadening at cryogenic temperatures. We, therefore, introduce a sigma in the form of

$$\sigma = \max(\sigma_c\hbar\omega_i, k_BT) \tag{22}$$

where σ_c is a constant.

Fig. 3 shows the variation of σ_c between $1/5$ and 3. For this work, $\sigma_c = 1/3$ was chosen because, then, in a range of 3σ at low temperature, each vibrational function $\eta_{i\alpha}$ can interact with at least one $\eta_{j\beta}$.

B. Temperature Dependence

One main feature of the full quantum mechanical model compared to the classical model is its validity at low temperatures [10], [19].

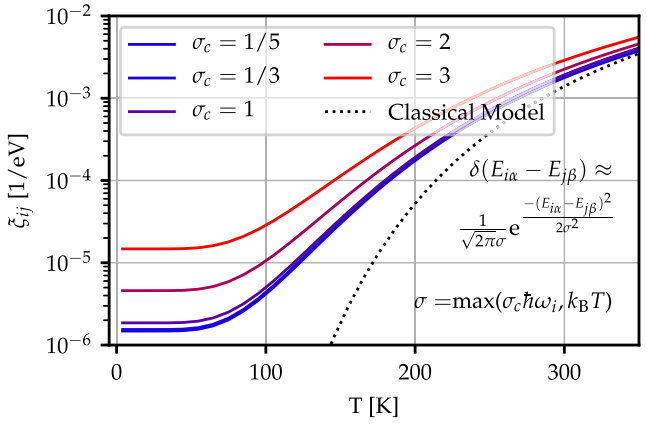

Fig. 3. For comparing the full QM model and the WKB/SPM approximation, it is necessary to expand the Dirac delta distribution $\delta(E_{i\alpha} - E_{j\beta})$ in (2) to a Gaussian function. The choice of the width of the function is crucial. The impact of the minimal width $\sigma = \sigma_c(E_{i1} - E_{i0})$ is shown in the figure for different constants σ_c.

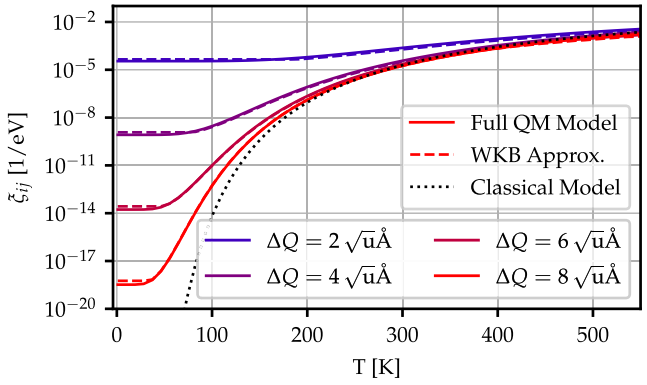

Fig. 4. Line shape function ξ_{ij} in the classical approximation (dotted black) is independent of the configurational coordinate ΔQ, while there is a strong ΔQ dependence in the full QM model and the WKB approximation using $E_R = 2.5$ eV, $R_{ij} = 1$, and $\Delta E = 0$ eV. Due to nuclear tunneling, for small ΔQ values, the line shape function does not freeze out toward low temperatures as the classical approximation does but stays constant instead.

This can be seen in Fig. 4, where the line shape function, which is proportional to the classical transition rate (dashed black line), freezes out for $T \to 0$ K, whereas the transition rates correctly remain constant below a certain temperature due to nuclear tunneling when using the full QM model. The transition rate within the QM model depends on the configurational coordinate, which was varied between 2 and 8 $\sqrt{\text{u}}$Å, a typical range for oxide defects obtained in density functional theory (DFT) calculations [19]. The WKB model can reproduce the convergence of the full QM model very precisely, as can be seen in the dashed lines in the plot.

The convergence to a constant line shape function at cryogenic temperatures can be expressed as an effective barrier lowering. This can be done by using the relation

$$E_{\text{eff}} = -k_B \frac{\partial \ln(\xi(T))}{\partial(1/T)} \tag{23}$$

to compute effective barriers from the corresponding line shape functions shown in Fig. 4. Note that, with this definition, the barrier of the classical model becomes slightly temperature dependent and deviates from ϵ_{ij} due to the temperature dependence of the carrier concentration and thermal velocity in (1).

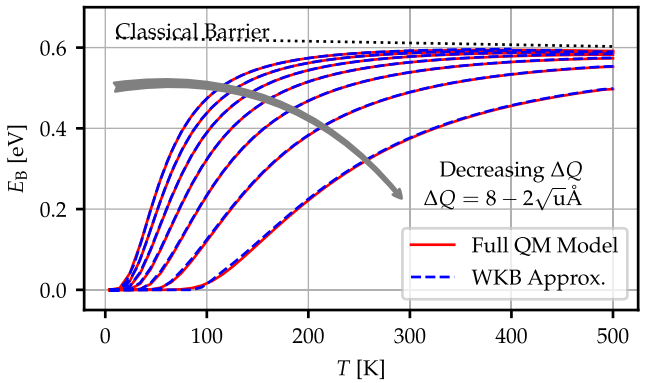

Fig. 5. Using the computed line shape functions in Fig. 4 an effective barrier for $\Delta Q = 2\ \sqrt{\text{u}}$Å to $8\ \sqrt{\text{u}}$Å can be calculated using (23). The classical approximation strongly overestimates the effective energy barrier for low temperatures.

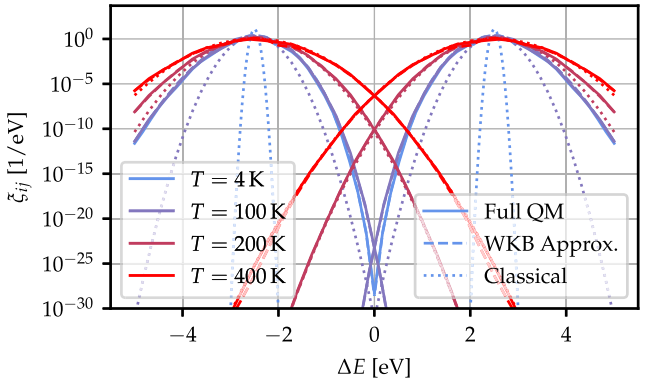

Fig. 6. Behavior of the line shape function ξ_{ij} (left) and ξ_{ji} (right) at $T =$ 4 K (blue), $T = 100$ K (violet), $T = 200$ K (magenta), and $T = 400$ K (red) is shown for different energetic displacements ΔE. While the classical line shape function freezes out except in a very small region with a barrier close to $\epsilon_{ij} = 0$ eV, the WKB line shape function and the full quantum mechanical line shape function deliver large values even at $T = 4$ K.

As can be seen in Fig. 5, at low temperatures, the effective barrier approaches 0 eV, which implies a temperature-independent rate according to the Arrhenius law. Toward high temperatures, the quantum mechanical transition rates approach the classical barrier. However, at 300 K, the effective barrier is essentially lower for small ΔQ, which implies an underestimation of the line shape function for certain defect configurations.

The behavior of the line shape function for different energy offsets is shown in Fig. 6 for the temperatures 400 K, 200 K, 100 K, and 4 K (red to blue). Even in the worst case, in the tail of the line shape function, the difference between the full QM model and the WKB approximation remains within one order of magnitude, which we consider reasonable for a quantity extending over 30 orders of magnitude, as can be seen in Fig. 6. At low temperatures, the classical barrier freezes out except for very small barriers on the order of $k_B T$, which corresponds to the maxima of the line shape functions in Fig. 6. The rates at 100 K and 4 K lie on top of each other for the displacement $\Delta Q = 2\ \sqrt{\text{u}}$Å, as can be expected from Fig. 4, where, for an energetic displacement $\Delta E = 0$ eV, the transition rates at both temperatures are identical.

TABLE II
SAMPLING PARAMETER SPACE

Parameter space			
Parameter	min.	max.	gridpoints
E_{R}[eV]	0.5	6	10
ΔE [eV]	-3	3	10
ΔQ [$\sqrt{\mathrm{u}}$Å]	1	6	10
R_{ij} [1]	0.7	1.3	3
T [K]	4	500	10

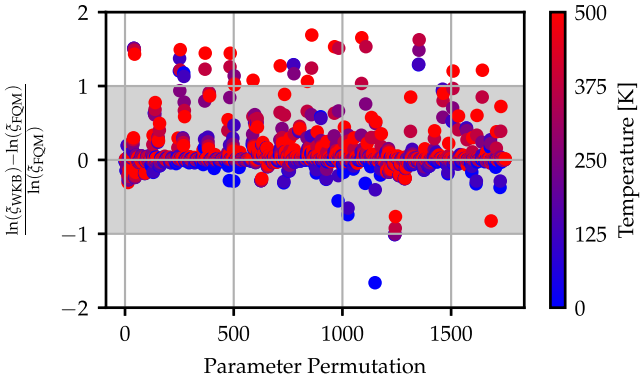

Fig. 7. Huang–Rhys parameters E_{R}, ΔE, R_{ij}, and ΔQ and the temperature are varied, as shown in Table II. For each variation, the relative error (24) is computed and plotted against the permutation number. The relative error of more than 97% of the permutations is in the gray interval $[-1, 1]$.

C. Relative Error of the WKB Approximation

For a comparison of the full quantum mechanical model and the WKB approximation, the parameters E_{R}, R_{ij}, ΔQ, ΔE, and T, are varied, as shown in Table II. For comparing the WKB approximation and the full QM model, the relative error of the logarithmic line shape functions is computed because the rates vary over at least 30 orders of magnitude

$$\delta_{\mathrm{rel}} = \frac{\ln(\xi_{\mathrm{WKB}}) - \ln(\xi_{\mathrm{FQM}})}{\ln(\xi_{\mathrm{FQM}})}. \tag{24}$$

For more than 97% of the parameter variations within the defined parameter space, the relative errors lie in the interval $[-1, 1]$. This can be seen in Fig. 7 for different temperatures. The largest relative errors arise from large positive ΔE. In this parameter range, both the classical barrier and the quantum mechanical barriers are very high, which leads to very low transition rates. Thus, we consider an error of one to two orders of magnitude in this range as acceptable. Furthermore, at cryogenic temperatures, transition rates with a positive ΔE do not contribute because only the ground state is occupied, as can be seen in Fig. 2.

IV. BENCHMARK

In device simulations, typically thousands of defects have to be considered, and the transition rate of each defect needs to be computed for every timestep; therefore, a computationally efficient charge transition model is necessary. For a given bias condition, the time to compute the transition rates in the classical model is independent of the relative position of the two potential energy surfaces (PECs).

The full quantum mechanical summation (2) on the other hand depends on the relative position of the PECs in several ways. First, it is necessary to compute all occurring overlap terms (5); this can be done efficiently with the recurrence relations derived by Schmidt [24]. However, computing the overlap matrix still scales with $O(n^2)$, where n is the index of the highest occurring vibrational wave function. The summation of the overlaps multiplied with the occupancy also scales with n though the scaling of the full quantum mechanical transition rate is dominated by the scaling of the overlap matrix. This can be seen in Fig. 8, where the size of the overlap matrix scales in the same way as the computation time of the full quantum mechanical model (blue).

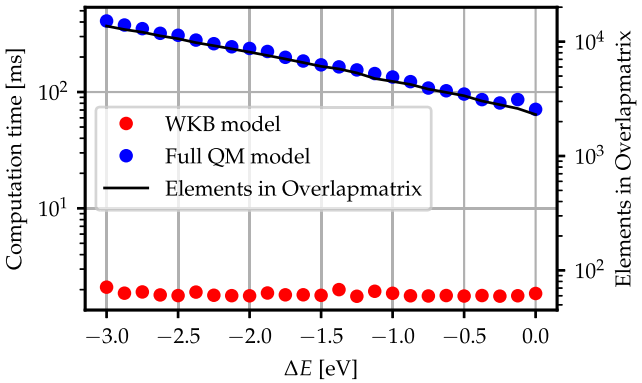

Fig. 8. Computation time using the full quantum mechanical transition rate scales quadratically with the number of overlapping eigenstates (and, therefore, with the elements in the overlapmatrix shown in black) and, hence, with energetic distance ΔE. Using the WKB approximation, the time limiting step is finding the extremum E^*, which is independent of the number of eigenstates.

The computation of the transition rate using the WKB approximation (see Fig. 8, in red), on the other hand, is independent of the relative position of the PECs because it is only necessary to compute the extrema E^* and Q_c and perform some algebraic operations with them. Thus, the calculation time of the transition rate is constant, and for $\Delta E = 0$ eV, the computation is approximately faster by a factor of 100 using the parameters given in the figure caption.

V. CONCLUSION

For the computation of NMP charge transition rates at cryogenic temperatures, it is necessary to use a full quantum mechanical description based on Fermi's golden rule. Since such calculations are computationally expensive and, thus, not feasible for efficient TCAD modeling, we propose a computationally efficient WKB approximation. We benchmark the approximation with the full quantum mechanical model and show that the approximation delivers excellent results for a wide range of input parameters.

REFERENCES

[1] I. D. C. Lamb *et al.*, "A FPGA-based instrumentation platform for use at deep cryogenic temperatures," *Rev. Sci. Instrum.*, vol. 87, no. 1, pp. 1–8, 2016, doi: 10.1063/1.4939094.
[2] J. M. Hornibrook *et al.*, "Cryogenic control architecture for large-scale quantum computing," *Phys. Rev. A, Gen. Phys.*, vol. 3, no. 2, pp. 1–9, Feb. 2015, doi: 10.1103/PhysRevApplied.3.024010.
[3] D. J. Reilly, "Engineering the quantum-classical interface of solid-state qubits," *npj Quantum Inf.*, vol. 1, no. 1, pp. 1–10, Dec. 2015, doi: 10.1038/npjqi.2015.11.

[4] H. Homulle *et al.*, "A reconfigurable cryogenic platform for the classical control of quantum processors," *Rev. Sci. Instrum.*, vol. 88, no. 4, pp. 343–346, 2017, doi: 10.1063/1.4979611.

[5] D. J. Reilly, "Challenges in scaling-up the control interface of a quantum computer," in *Tech. Dig.*, Dec. 2019, p. 31, doi: 10.1109/IEDM19573.2019.8993497.

[6] A. Beckers, F. Jazaeri, and C. Enz, "Theoretical limit of low temperature subthreshold swing in field-effect transistors," *IEEE Electron Device Lett.*, vol. 41, no. 2, pp. 276–279, Feb. 2020, doi: 10.1109/LED.2019.2963379.

[7] A. Beckers, F. Jazaeri, A. Grill, S. Narasimhamoorthy, B. Parvais, and C. Enz, "Physical model of low-temperature to cryogenic threshold voltage in MOSFETs," *IEEE J. Electron Devices Soc.*, vol. 8, pp. 780–788, 2020, doi: 10.1109/JEDS.2020.2989629.

[8] A. Beckers, F. Jazaeri, and C. Enz, "Characterization and modeling of 28-nm bulk CMOS technology down to 4.2 K," *IEEE J. Electron Devices Soc.*, vol. 6, pp. 1007–1018, 2018, doi: 10.1109/JEDS.2018.2817458.

[9] A. Grill *et al.*, "Reliability and variability of advanced CMOS devices at cryogenic temperatures," in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Apr. 2020, pp. 1–7, doi: 10.1109/IRPS45951.2020.9128316.

[10] J. Michl *et al.*, "Quantum mechanical charge trap modeling to explain BTI at cryogenic temperatures," in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Apr. 2020, pp. 1–6, doi: 10.1109/IRPS45951.2020.9128349.

[11] J. H. Scofield, N. Borland, and D. Fleetwood, "Temperature-independent switching rates for a random telegraph signal in a silicon metal–oxide–semiconductor field-effect transistor at low temperatures," *Appl. Phys. Lett.*, vol. 76, no. 22, pp. 3248–3250, 2000.

[12] T. Knobloch *et al.*, "Analysis of single electron traps in nano-scaled MoS_2 FETs at cryogenic temperatures," in *Proc. Device Res. Conf. (DRC)*, Columbus, OH, USA, 2020, pp. 52–53.

[13] T. Grasser, "Stochastic charge trapping in oxides: From random telegraph noise to bias temperature instabilities," *Microelectron. Rel.*, vol. 52, no. 1, pp. 39–70, Jan. 2012, doi: 10.1016/j.microrel.2011.09.002.

[14] A.-M. El-Sayed, M. B. Watkins, V. V. Afanas'ev, and A. L. Shluger, "Nature of intrinsic and extrinsic electron trapping in SiO_2," *Phys. Rev. B, Condens. Matter*, vol. 89, Mar. 2014, Art. no. 125201, doi: 10.1103/PhysRevB.89.125201.

[15] J. Strand, M. Kaviani, D. Gao, A.-M. El-Sayed, V. V. Afanas'ev, and A. L. Shluger, "Intrinsic charge trapping in amorphous oxide films: Status and challenges," *J. Phys., Condens. Matter*, vol. 30, no. 23, May 2018, Art. no. 233001, doi: 10.1088/1361-648x/aac005.

[16] A.-M. El-Sayed, Y. Wimmer, W. Goes, T. Grasser, V. V. Afanas'ev, and A. L. Shluger, "Theoretical models of hydrogen-induced defects in amorphous silicon dioxide," *Phys. Rev. B, Condens. Matter*, vol. 92, no. 1, p. 14, Jul. 2015, doi: 10.1103/PhysRevB.92.014107.

[17] Y. Wimmer, A.-M. El-Sayed, W. Gös, T. Grasser, and A. L. Shluger, "Role of hydrogen in volatile behaviour of defects in SiO_2-based electronic devices," *Proc. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 472, no. 2190, pp. 1–23, 2016, doi: 10.1098/rspa.2016.0009.

[18] S. D. Ganichev, I. N. Yassievich, V. I. Perel, H. Ketterl, and W. Prettl, "Tunneling ionization of deep centers in high-frequency electric fields," *Phys. Rev. B, Condens. Matter*, vol. 65, no. 8, Feb. 2002, Art. no. 085203, doi: 10.1103/PhysRevB.65.085203.

[19] W. Goes *et al.*, "Identification of oxide defects in semiconductor devices: A systematic approach linking DFT to rate equations and experimental evidence," *Microelectron. Rel.*, vol. 87, pp. 286–320, Aug. 2018, doi: 10.1016/j.microrel.2017.12.021.

[20] G. Rzepa *et al.*, "Efficient physical defect model applied to PBTI in high-κ stacks," in *Proc. IEEE Int. Rel. Phys. Symp.*, Apr. 2017, p. XT-11, doi: 10.1109/IRPS.2017.7936425.

[21] W. Nolting, *Grundkurs Theoretische Physik 5/2: Quantenmechanik-Methoden und Anwendungen* (Springer-Lehrbuch). Berlin, Germany: Springe, 2011.

[22] T. Tewksbury, "Relaxation effects in MOS devices due to tunnel exchange with near-interface oxide traps," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, USA, 1992. [Online]. Available: http://dspace.mit.edu/handle/1721.1/13238

[23] K. Huang and A. Rhys, "Theory of light absorption and non-radiative transitions in F-centres," in *Selected Papers of Kun Huang: With Commentary*, vol. 4, 2000, pp. 74–92, doi: 10.1142/9789812793720_0007.

[24] P. Schmidt, "Computationally efficient recurrence relations for one-dimensional Franck–Condon overlap integrals," *Mol. Phys.*, vol. 108, no. 11, pp. 1513–1529, 2010, doi: 10.1080/00268971003762142.

[25] T. Holstein, "Quantal occurrence-probability treatment of small-polaron hopping," *Phil. Mag. B*, vol. 37, no. 1, pp. 49–62, Jan. 1978, doi: 10.1080/13642817808245306.

[26] T. Markvart, "Determination of potential surfaces from multiphonon transition rates," *J. Phys. C, Solid State Phys.*, vol. 14, no. 15, pp. L435–L440, May 1981, doi: 10.1088/0022-3719/14/15/002.

[27] T. Markvart, "Multiphonon transitions between adiabatic potential curves," *J. Phys. C, Solid State Phys.*, vol. 17, no. 35, pp. 6303–6316, Dec. 1984, doi: 10.1088/0022-3719/17/35/006.

[28] J. Mathews and R. Walker, *Mathematical Methods of Physics*, W. A. Benjamin, Ed. Reading, MA, USA: Addison-Wesley, 1970.

[29] A. Nitzan, *Chemical Dynamics in Condensed Phases: Relaxation, Transfer, and Reactions in Condensed Molecular Systems* (Oxford Graduate Texts). Oxford, U.K.: Oxford Univ. Press, 2013.