IOPscience iopscience.iop.org

Home Search Collections Journals About Contact us My IOPscience

Full-band cellular Monte Carlo simulations of terahertz high electron mobility transistors

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

2008 J. Phys.: Condens. Matter 20 384201

(http://iopscience.iop.org/0953-8984/20/38/384201)

View the table of contents for this issue, or go to the journal homepage for more

Download details:
IP Address: 129.252.86.83
The article was downloaded on 29/05/2010 at 15:06

Please note that terms and conditions apply.

IOP Publishing — Journal of Physics: Condensed Matter
J. Phys.: Condens. Matter **20** (2008) 384201 (9pp) doi:10.1088/0953-8984/20/38/384201

Full-band cellular Monte Carlo simulations of terahertz high electron mobility transistors

R Akis, J S Ayubi-Moak[1], D K Ferry, S M Goodnick, N Faralli and M Saraniti

Department of Electrical Engineering and Arizona Institute for Nano-Electronics, Arizona State University, Tempe, AZ 85287-5706, USA

Received 10 March 2008, in final form 21 April 2008
Published 21 August 2008
Online at stacks.iop.org/JPhysCM/20/384201

Abstract
High electron mobility transistors (HEMTs) have become important for high frequency and low noise applications. There are devices now operating with a cutoff frequency, f_T, of several 100 GHz. Through simulation, we have been investigating how these frequencies may be pushed even higher, and have found that it may be possible to achieve an f_T of over 3 THz. For this, we have used a full-band, cellular Monte Carlo transport program, coupled to a full Poisson solver, to study a variety of InAs-rich, InGaAs pseudomorphic HEMTs and their response at high frequency, concentrating on devices with a structure (from the substrate) InP/InAlAs/InGaAs/InAlAs/InGaAs, with the quantum well composed of $In_{0.75}Ga_{0.25}As$. We have studied gate lengths over the range 10–70 nm and various source–drain spacings. The performance of scaled devices has been evaluated to determine the ultimate frequency limit. Here, the importance of the effective gate length has been evaluated from the properties internal to the device.

(Some figures in this article are in colour only in the electronic version)

1. Introduction

In the search for sources and detectors in the terahertz regime, the high electron mobility transistor (HEMT) has become one of the devices of choice [1]. Experimentally, the use of In-rich InGaAs on InP-based pseudomorphic high electron mobility transistors (PHEMTs) has recently resulted in submillimeter-wave amplifiers above 300 GHz [2] and individual transistor cutoff frequencies approaching 600 GHz [3]. Improved figures of merit, such as cutoff frequency, f_T, maximum frequency of oscillation, f_{max}, and maximum gain in InP-based HEMTs are primarily a result of higher carrier mobilities and correspondingly larger peak carrier velocities achieved through the strained growth of InGaAs/InAlAs on InP and better separation between conduction carriers and ionized dopant atoms. In addition, the large conduction band offset between InGaAs and InAlAs (0.53–0.63 eV) [4] provides enhanced confinement of carriers in the resulting two-dimensional electron gas (2DEG) allowing for higher sheet carrier densities (2×10^{12}–4×10^{12} cm^{-2}) within the active channel region of the device.

Over the last few years, we have made a major effort to simulate such devices with the aim of optimizing their performance. Amongst our results, we have found that f_T may be greatly enhanced by reducing the gate length and the source–drain spacing (SDS), with f_T s above 1.5 THz obtained for a device with 20 nm gate length and a 500 nm SDS [5]. More recently, we showed that using a very short device (300 nm SDS) and scaling the gate length was successful in establishing a theoretical upper limit for f_T in an InGaAs PHEMT [6, 7]. Through an appropriate definition of an effective gate length and extrapolation to vanishingly small physical gate length, we found that this is of the order of 3 THz, a result far higher than some previous estimates of this quantity. In establishing such limits, we found that using the results of a full RF analysis [6, 7] is crucial in formulating a definition of an effective gate length. Using the depletion length as the effective gate length [8], and using that to estimate f_T, can lead to an underestimate of the theoretical value of this quantity by

[1] Present address: Department of Electronics and Electrical Engineering, Glasgow University, Glasgow G12 8LT, UK.

 © 2008 IOP Publishing Ltd Printed in the UK

more than a factor of two, as our more mathematical analysis has shown.

In this paper, we will review our recent work and discuss the current understanding that we have of the physics that is important for ultrashort gate length HEMTs, using the In-rich InGaAs channel HEMT as our prototypical device. Here, we will discuss the role of series resistances, results on scaling the gate length, and the importance of an appropriate definition of the effective gate length. The importance of the contact resistance, source–drain spacing, gate–channel spacing, and quantum well width will all be discussed.

The paper is organized as follows. In section 2, we shall briefly describe the full-band, cellular Monte Carlo simulation tool that has been developed by our group. We shall also discuss the device structure that is used in the simulation, and give some details of the band structure and phonon models that have been utilized in obtaining our results. In section 3, we will discuss the role of resistance in the devices and the effect of varying the total device length. In section 4, we shall discuss the results of scaling the gate length and how one may infer an upper limit for the cutoff frequency. We draw some conclusions in the final section.

2. The full-band Monte Carlo simulation approach as applied to the PHEMT device structure

The full-band simulator used in this work is based upon a hybrid transport approach discussed previously [9]. This simulator combines a traditional ensemble Monte Carlo method (for low carrier energies) with a cellular Monte Carlo approach (at high carrier energies). In our simulator, the solution of Poisson's equation is computed using a fast multi-grid technique [10], which solves the full set of electrostatic field equations resulting from a center-difference discretization of the 3D Poisson equation over a set of grids with varying coarseness. These grids act in concert to simultaneously reduce the different low and high frequency components of the error, resulting in faster and more robust convergence [11].

A full-band representation of the electron dispersion relationship is computed via the empirical pseudopotential method (EPM) [12] and includes local, non-local, and spin–orbit interactions in the calculation. Non-local corrections are included only for the off-diagonal Hamiltonian elements, in keeping with the original formulations of this effect [13, 14]. The pseudopotential parameters that we used are adjusted to give a best fit to the valence band at Γ and the positions of the three conduction band minima, rather than the overall optical absorption properties, although the starting parameters were taken from Chelikowsky and Cohen [12] and from Pötz and Vogl [4].

An important effect that must be accounted for in our simulations is the strain caused by the lattice mismatch at the $In_{0.52}Al_{0.48}As/In_{0.75}Ga_{0.25}As$ heterojunctions. This strain is the result of lattice deformation due to growth layer accommodation of the lattice spacing to the underlying substrate material. These physical changes in the crystal lattice result in changes in the Brillouin zone (BZ) and a subsequent breaking of degeneracy in the electronic band structure as the material changes from a zinc-blende structure due to a tetragonal distortion. The end result is an enlarged irreducible BZ wedge that has different symmetry properties than that of the unstrained zinc-blende structure. In order to model the effects of the strained band structure in our simulator, we matched the lattice constant of $In_{0.75}Ga_{0.25}As$ to that of the InP substrate and then systematically adjusted and fine-tuned the correct symmetric and anti-symmetric pseudopotential parameters to match as closely as possible a band gap energy of 0.58 eV [15], and the corresponding energy offsets between Γ–L and Γ–X valleys in the electron dispersion relationship. Importantly, in our simulations, we maintained a zinc-blende material structure and simply applied the necessary strain *hydrostatically*. This approximation assumes an isotropically strained unit cell and allows us to utilize a smaller irreducible wedge than would be necessary for a tetragonal crystal structure. We took this approach for the simple reason that the nature of the tetragonally distorted cell and its band structure is not known at this time. This is further complicated by the fact that Mikkelsen and Boyce [16] have shown that the InAs and GaAs nearest-neighbor bond lengths remain nearly constant at the binary values over the entire composition range. How this varies with the strain and the tetragonal distortion is currently not known. In view of this, approximating the band structure with the hydrostatic strain model is probably as good an approximation as any available at the current time.

Phonons in ternary compounds often exhibit two-mode behavior, resulting in two sets of optical branches instead of one. The use of InGaAs in particular results in both GaAs-like and InAs-like modes for which the compositional dependences of longitudinal and transverse optical modes near the center of the Brillouin zone have been measured experimentally by means of Raman scattering and compared with theoretical calculations [17]. In order to determine the proper electron–phonon scattering rates, the full phonon spectrum for the material is computed over the first BZ in our simulator. We accomplish this by using a 14-parameter valence shell model [18–20] to calculate the lattice energy. However, the LO modes are affected by the dielectric properties as well as the elastic properties, and it is not known how these vary with the strain in the crystal, or with the tetragonal distortion. As a result, for the results presented in the following sections, phonon scattering rates were calculated using only the InAs phonon modes. This approximation was used due to the In-rich nature of the simulated alloy, and the fact that the InAs-like LO mode would dominate the overall scattering. This was found to yield device simulation results comparable with measured device characteristics [5].

Our code includes scattering mechanisms due to deformation potential, optical and acoustic phonons, and polar optical phonons. Impurity scattering is included via the Ridley model [21]. All of the possible initial and final momentum states for each possible scattering mechanism are pretabulated and stored in extensive data tables which are subsequently loaded into random access memory (RAM) during simulation runtime. This preprocessing step simplifies the final state selection process to the generation of a single random number, significantly reducing the time required to update the corresponding carrier energy and momentum [9].

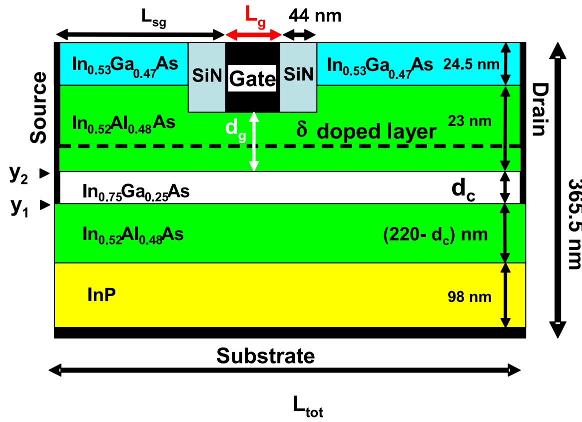

Figure 1. Simulated two-dimensional PHEMT structure. For the purposes of clarity, the figure is not to scale. The various layers are nominally undoped, but we have used a value of 10^{12} cm^{-3} in the code to ease convergence of the Poisson solver. The layer thicknesses shown here apply only to the 300 nm device.

While our simulations are entirely particle based, we do compensate somewhat for quantum effects in the channel, specifically, the charge setback from the gate that one expects due to the effects of subband quantization. This has been accounted for in our simulations using an effective potential model [22].

The basic device structure used in our 2D simulations is shown in figure 1. The computational layout is a simplified version of a recessed T-gate structure [2]. This simplification has been made to ease the simulation task. While the structure of [2] appears to have SiN only on the surfaces of the semiconductor and the gate, we have filled the entire (square) recess with this insulator. The thicknesses and doping concentrations of the heterolayers were chosen to closely match those actual devices. Source and drain electrodes are treated as vertical ohmic contacts extending down from the cap to the active channel region. The gate electrode is treated as an absorbing Schottky contact with a 0.8 eV barrier. Conduction band offsets of 0.53 eV were used for both the $In_{0.53}Ga_{0.47}As/In_{0.52}Al_{0.48}As$ and $In_{0.75}Ga_{0.25}As/In_{0.52}Al_{0.48}As$ heterojunctions, with 0.34 eV used at the $In_{0.52}Al_{0.48}As/InP$ interface. Chosen to match experimental observation [2], a δ-doping layer concentration of 3.5×10^{12} cm^{-2} has been used.

In the remainder of this paper, results are shown for devices ranging in total length from 640 nm to 2 μm, while a second set of results were obtained with L_{tot} fixed at 300 nm in each case. For the longer devices, tensor product grid grids of 320 × 95 up to 1000 × 95 were used with uniform grid spacings in the x direction ($\Delta x = 2$ nm) and non-uniform Δy spacings, between 0.5 and 14 nm. A finer grid is used across the active channel region, the highly doped delta-doping layer and cap layers, and a coarser grid utilized in the buffer and substrate layers. The δ-doping layer was spread over three grid cells along the positive y direction, 4 nm from the channel. It was found that spreading the δ-doping over a number of grid cells allowed for a better fit to the threshold voltage and the sheet charge density when comparisons to real devices were made [5]. For the $L_{\mathrm{tot}} = 300$ nm results shown in section 4, the system was represented on a 150 × 148 grid. The number of grid points along the perpendicular direction is higher since more grid points with the shorter $\Delta y = 0.5$ spacing were included. This allowed for a finer control of the gate to channel distance and the channel thickness for our scaling simulations. In addition, the δ-doping layer was spread over two grid cells along the positive y direction, 1 nm from the channel for the 300 nm device.

In all cases, the entire structure sits atop an InP substrate and an 'undoped' $In_{0.52}Al_{0.48}As$ buffer layer is added next. During the course of our studies, we have varied the thickness of this buffer layer and even taken it to the bottom of the simulation domain without producing any significant change in the results. We have left it at the nominal thickness of ~200 nm here (when the channel thickness, d_c, is reduced from 18 to 10 nm in section 4, we make this buffer layer thicker to ensure that the total height of the device is constant; all results in section 3 use $d_c = 18$ nm). The InP substrate is included as we know the donor levels here better. While we presume that the layers are nominally undoped (or, more properly, not intentionally doped), we have assumed a doping of 1×10^{12} cm^{-3} as the background doping for all layers, other than the δ-doped layer, in order to ease the convergence of the Poisson solver.

The back contact of the substrate is treated as a constant potential surface, with a local potential relative to the Fermi energy (which is the reference level for all potentials in the simulation) that is set by the doping of the substrate. The rationale for this lies in submillimeter-wave device packaging. Normally, these devices are mounted on a metalized carrier. In addition, the source itself is well grounded to assure good signal grounds. As a result of this, we use a vertical source contact (on the left in figure 1) and similarly with the drain contact. But, the metal carrier also provides excellent screening of the substrate so that there is no charge build-up in the substrate. Hence, we find it acceptable to treat the substrate as the equipotential surface described here.

3. Results as a function of device length and the role of contact resistance

One of the most important issues in trying to improve device performance is examining and understanding the role of contact resistance. In our simulator, we can control this property by adjusting the doping concentration in the small region which is defined to be the contact. We have found that using too low a value of the doping leads to potential drops in the contact regions, and subsequent poorer device performance. Moreover, we have established that the actual simulated contact resistance can be related to observed experimental values [5], so that good agreement with experimental I_d–V_d characteristics can be obtained. More important, however, is the fact that the part of the device between the source contact and the gate also provides a significant series resistance, which affects the device performance. Thus, the simulated device characteristics are related to the manner in which device contacts are handled.

Here, both the source and drain contacts are treated as ohmic contacts while the gate electrode is treated as an

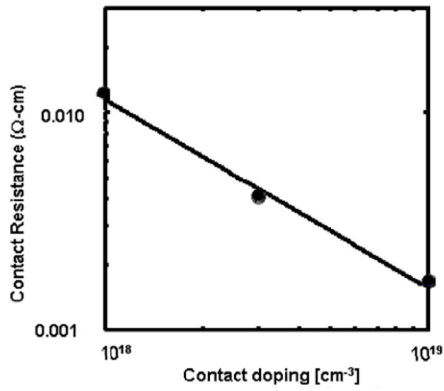

Figure 2. Variation of the effective source resistance created by the potential drop adjacent to the source contact, as a function of the contact doping. The source resistance is estimated from the reduction in transconductance with resistance, as shown in equation (1).

absorbing Schottky contact with a barrier height of 0.8 eV. In the context of our simulations, 'ohmic' contact means the following: the 'effect' that a metallic contact has on the semiconductor material directly adjacent to it is what is actually being simulated during runtime. Two primary boundary conditions are established on these 'contact cells' within the simulation domain. First, a fixed potential or enforced Dirichlet boundary condition (necessary for the solution of Poisson's equation) is imposed. The second condition is that of charge neutrality. This boundary condition is imposed on the Boltzmann transport equation, which is solved stochastically by the cellular Monte Carlo portion of the code. Charge neutrality is assured by assigning a doping concentration to each simulated contact cell during the initialization phase of the simulation. A fixed number of simulated particles are then maintained in each contact cell throughout the total simulation time and carriers are subsequently injected and/or ejected following each periodic update of the electrostatic potential ensuring charge neutrality within these regions [23]. The simulated doping within the 'contact cells' creates a potential drop near the ohmic contacts that emulates an *effective* source resistance. Thus, by varying the doping concentration within the contact cells, we can easily study the effect of an *internal* contact resistance in our model. This is illustrated in figure 2. Increasing the doping density in the contact regions means that more carriers are initialized in these regions and thus a larger integer number of carriers are maintained via charge neutrality in the contact cells. The effect of increasing the doping density in the regions is clearly seen in the effective resistance curve. In this case, a set of 70 nm gate simulations were performed over a range of contact dopings. The effective resistance can be determined from the reduction in transconductance from that of nearly infinite contact doping, using the common relationship

$$g_m = \frac{g_{m0}}{1 + g_{m0} R_C}. \tag{1}$$

If the doping is set too low, then charge neutrality cannot be maintained in the contact regions, so only the range of dopings where we have confidence in the results is shown. A similar

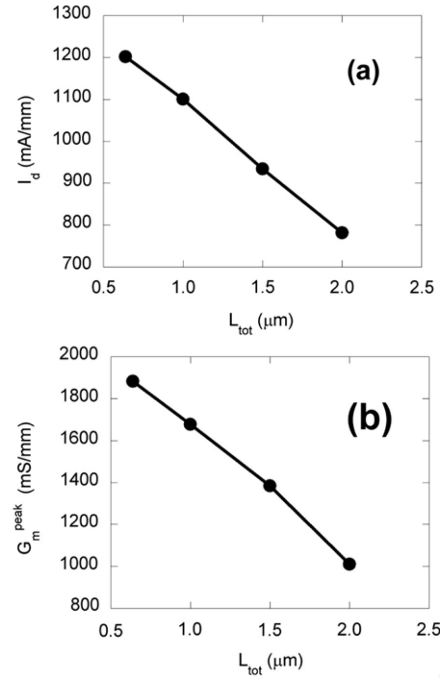

Figure 3. (a) Calculated output current of a 70 nm gate PHEMT as a function of source–drain spacing, for a δ-layer doping concentration of 3.5×10^{12} cm^{-2} and a simulated 'contact' doping of 3×10^{18} cm^{-3}. Here, $V_D = 1.0$ V and $V_G = 0.4$ V. (b) As in (a), but now the peak transconductance is plotted with $V_D = 1.0$ V.

result was obtained by actually measuring the potential drops at the contacts that were found from the Poisson solver. While the quantitative results differed, the trends were quite similar. Thus, the contacts are a crucial point in the simulation, and must be carefully addressed.

In a similar manner, the semiconductor material between the source and drain also provides a series resistance within the device. We have also studied this effect, particularly as the devices that we typically simulate are generally shorter than the experimental ones. In this case, a set of 70 nm gate simulations were performed over a range of device lengths from 0.64 to 2.0 μm. In each of these simulations the contact doping was fixed at 3×10^{18} cm^{-3}. In figures 3(a) and (b), we plot the drain current at $V_D = 1.0$ V and $V_G = 0.4$ V and the peak transconductance at $V_D = 1.0$ V. The peak transconductance has increased from 1500 mS mm^{-1} to almost 1900 mS mm^{-1}, as the channel length is shortened, indicating faster overall device response. An increase in the source–drain spacing reveals a decrease in the overall drive current and lower peak transconductance as L_{tot} is increased. This effect is best understood by relating the additional distance carriers must travel across the active channel region to an increase in the *effective* internal source to drain series resistance of the device. Thus, as L_{tot} is increased, this resistance increases and results in lower drive currents and reduced switching speeds.

In addition to static DC characterization, small-signal analysis was also performed in order to investigate the frequency response of each structure and to determine the corresponding maximum current gain achievable in each case. This analysis is performed by first biasing each device at

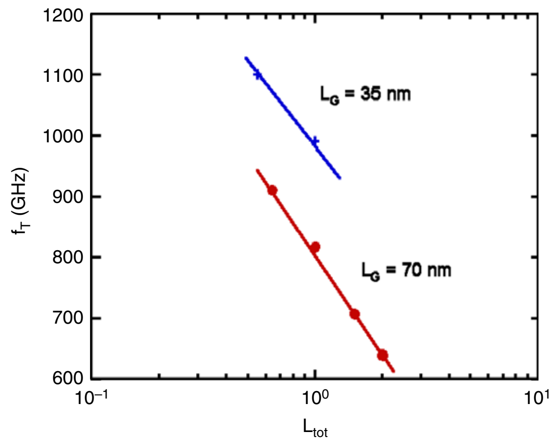

Figure 4. Variation of the cutoff frequency with the source–drain spacings for 35 and 70 nm gate lengths.

a known bias point in the operating region and allowing it to reach a quasi-static dc (steady-state) condition. Next, a small voltage (typically 200 mV) step is applied directly to the gate (drain) contact while holding the drain (gate) at a constant electrostatic potential. The corresponding gate (drain) currents are then stored after each update of the field equations and used during post-simulation analysis to compute the y-parameters [1] (via the Fourier transform) which are then used to extract the corresponding small-signal gains. In figure 4, we plot the variation of this cutoff frequency as determined from the current gain for the various device lengths, in order to see how it affects the high frequency performance. Here we consider gates of lengths of 70 and 35 nm. For the 70 nm gate length, f_T reveals an exponential increase in the cutoff frequency as the total device length, L_{tot}, is scaled downward (note the logarithmic scale). The 35 nm results appear to follow basically the same trend.

We now consider the 300 nm device. In figure 5(a), we map the position of the carriers, located between the source and the gate, in the full BZ for a simulation obtained in that case. Here, it may be observed that these carriers are almost entirely low energy carriers in the Γ valley of the conduction band. Hence, these carriers are moving relatively slowly and lead to the effective series resistance. A second important aspect is that the carriers begin to transfer to the L valleys as soon as they are accelerated under the gate. Once in the satellite valleys (either L or X), they remain in these valleys for a very long time. Because the mass in the Γ valley is so low, the density of states is also low and the resulting rate of scattering from L (or X) to Γ is quite small [24]. We see this in figure 5(b), where we plot the position of the carriers, located in the channel between the gate and the drain, in the BZ. Some carriers remain in the Γ valley, since the field and the velocity both drop in this region (beyond the gate). The peak of the velocity is actually reached when only a few per cent of the carriers have transferred to the satellite valleys. In this regard, since the fields are low in the gate–drain region, this is primarily a *drift* region with the carriers moving relatively slowly with the properties of the satellite valleys. However, one advantage is that the threshold energy for impact ionization is considerably higher in these

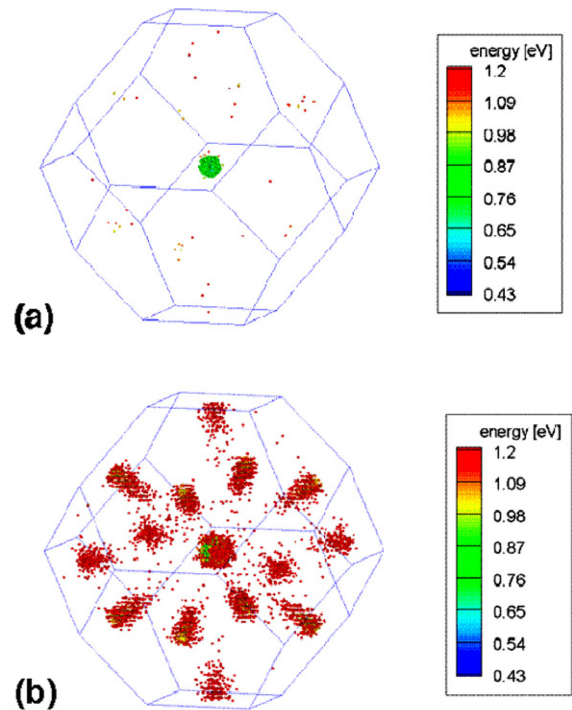

Figure 5. The projection of the simulated particles onto the Brillouin zone. Those particles located between the source and gate are depicted in (a), while (b) depicts those particles located between the gate and the drain. The shading is an energy code (which is measured relative to the valence band maximum). (c) Comparison of the transconductance of a 10 nm gate length, 10 nm channel device with a total source–drain spacing of 0.3 μm, as the source–gate distance is varied. Inset: cutoff frequency as a function of source–gate distance.

valleys ($\sim$1.5 × the L valley to valence band maximum gap), so it plays little role in device operation. In view of this, these devices can take high bias.

The effect that the slow moving carriers in the source to gate region have on device operation can be further demonstrated by looking at the effect that changing this distance has on the transconductance for a 10 nm gate length device. This is shown in figure 6, where we vary the source–gate distance over more than a factor of 2 (about the nominal distance). Here, it can be seen that lowering this distance from 140 to 60 nm produces almost a 40% increase in the transconductance. This change in distance also affects the gate–source capacitance. In the inset, we plot cutoff frequency over the same range of source–drain voltages. However, the latter only exhibits about a 10% increase. Nevertheless, it is clear that good design for mm-wave applications requires reduction of the device dimensions, particularly between the gate and the source.

4. Scaling the gate length to establish an upper limit for the cutoff frequency

In this section, we summarize the results that we have obtained by the scaling the gate length. The PHEMTs studied had gate lengths ranging 10 to 50 nm. In each of these simulations the contact doping was fixed at 3 × 10^{18} cm^{-3}. We have fixed

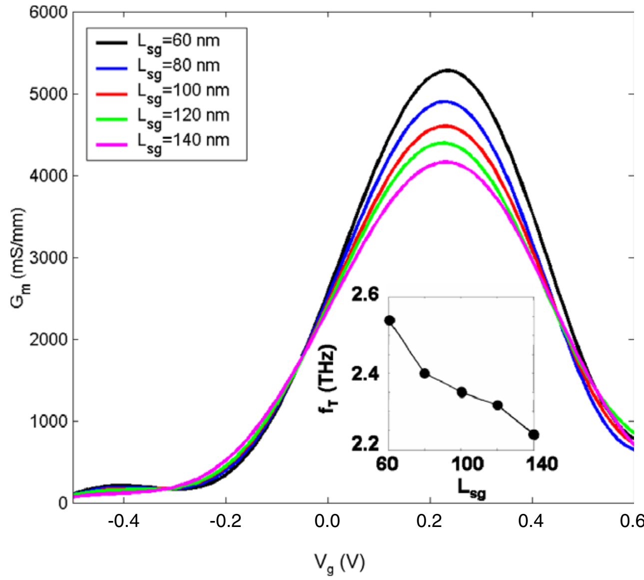

Figure 6. Comparison of the transconductance of a 10 nm gate length, 10 nm channel device with a total source–drain spacing of 0.3 μm, as the source–gate distance is varied. Inset: cutoff frequency as a function of source–gate distance.

$L_{\text{tot}} = 0.3$ μm, but the gate to channel separation is scaled with the gate length according to $d_g = L_g/5$. In figure 7(a), we plot the velocity along the channel as a function of position within the channel. Here, a channel thickness of $d_c = 18$ nm was employed. These velocities were computed by doing a weighted averaging along the y direction in the total channel region:

$$v_x^{\text{avg}} = \left(\int_{y_1}^{y_2} n v_x \, dy\right) \Big/ \left(\int_{y_1}^{y_2} n \, dy\right), \qquad (2)$$

where n is the electron density and y_1 and y_2 are the upper and lower boundaries of the channel as indicated in figure 1. It should be noted that the velocity begins to rise *before* the gate is reached, a result of field spreading in the gate direction [25]. The peak of the velocity, in each case, occurs before the end of the gate is reached, and this is the point at which significant numbers of carriers are being transferred to the satellite valleys, as mentioned above. Figure 7(b), we plot the transconductance as a function of gate voltage for these same devices. It may be seen that the transconductance is enhanced as the gate length is reduced, which implies a better frequency response as well.

In figure 8, we plot simulated drain current, I_d as a function of both gate voltage, V_g, and V_{DS}, for a 10 nm gate length in (a) and a 50 nm length in (b). While the currents at the upper end of the scale are similar in all cases, $\sim$2500 mA mm^{-1}, it is evident that the shorter gate length device has a higher threshold voltage. As one might expect, a higher gate voltage is required to pinch off a device with shorter gate length, though it should be reiterated that as part of the scaling, the gate is closer to the channel in that case ($d_g = 2$ nm, compared to $d_g = 10$ nm for the 50 nm gate). Referring back to figure 7(b), the point of peak transconductance is

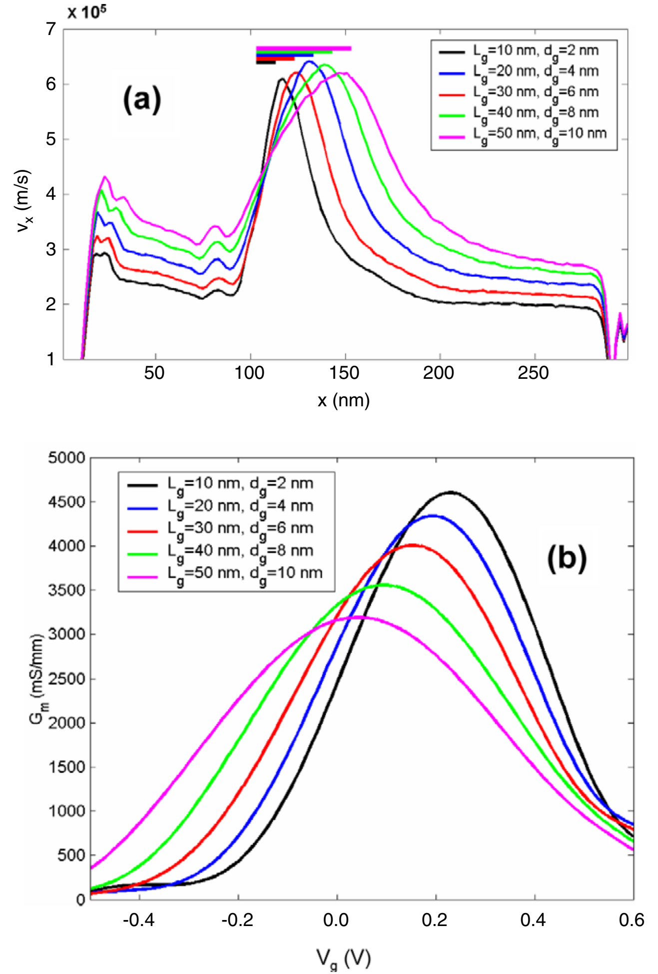

Figure 7. (a) The velocity along the channel for a set of scaled devices as discussed in the text. (b) The transconductance in these same devices for $V_{DS} = 1.0$ V.

shifting further to the right as the gate length is shortened, as one might expect from these I_d results.

In figures 9(a) and (b), frequency response results are shown for 10 and 50 nm gate lengths and a channel thickness $d_c = 18$ nm. A linear fit using a -20 dB/decade slope is also shown and indicates an f_T of 1.3 THz for the 50 nm gate length and 2.2 THz for a 10 nm gate length structure. Note that the 50 nm result for f_T is larger than either of the 35 nm results shown in figure 2(d). Thus, with a much shorter device (300 nm here compared with 640 nm for those earlier results), one can get away with using a longer gate length.

We shall now show how one can use these f_T results along with the results for velocity in the channel to formulate an appropriate definition for the *effective* gate length, L_g^{eff}, for these devices. As regards why such a quantity needs to be considered, it has been previously recognized in devices with very short gate lengths that the electrons in the channel behave as if they were under the influence of a gate longer than its actual physical dimensions [25]. Figure 10(a) shows the average velocity of the electrons in the channel along the x direction, as a function of the position along the channel for the 50 and 10 nm gate lengths. The cutoff frequency, f_T, is related

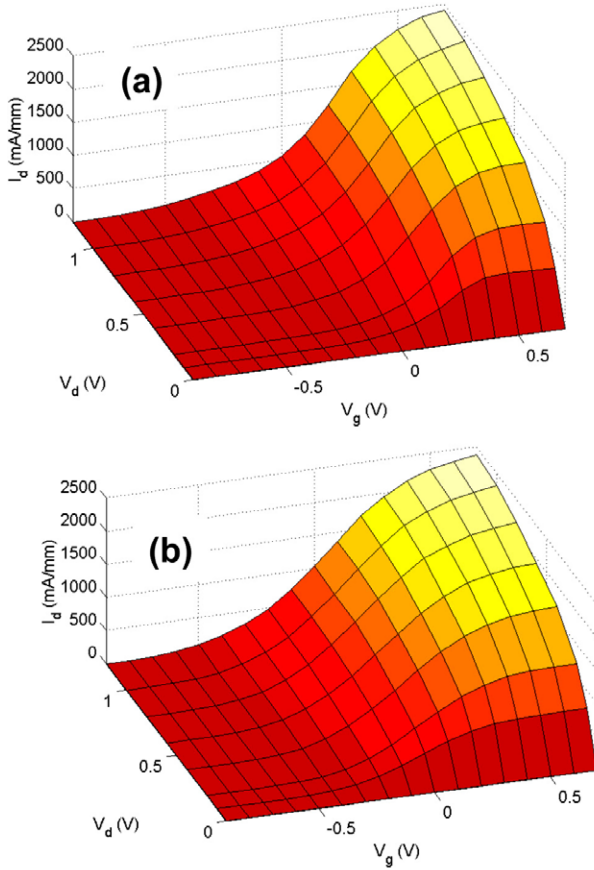

Figure 8. (a) The drain current as a function of both V_{DS} and V_g for a device with $L_g = 10$ nm and $d_c = 18$ nm. (b) As in (a), but for a device with a 50 nm gate length.

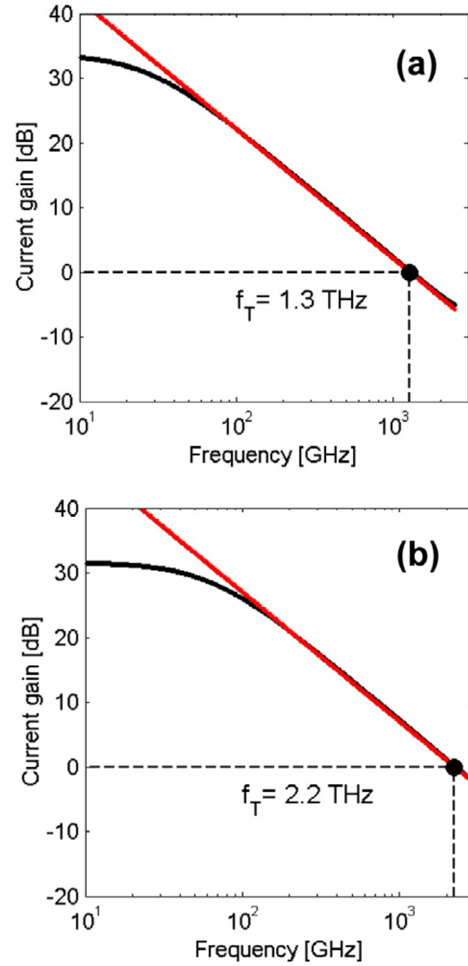

Figure 9. Frequency response characteristics for devices with $L_g = 50$ nm in (a) and $L_g = 10$ nm in (b). Here, $d_c = 18$ nm. (c) Average electron velocity along the channel with the actual and effective gate lengths also indicated (the gray and black dashed lines indicate the beginning and end points for the 10 and 50 nm gate length cases respectively). The dots represent the data points over which the integral for transit time was computed for each, as per equation (2). (d) The corresponding average electron density with estimated depletion lengths indicated for the same two cases.

to τ_T, the time of transit under the gate:

$$\tau_T = \frac{1}{2\pi f_T} = \int_0^{L_g^{\mathrm{eff}}} \frac{\mathrm{d}x'}{v_x^{\mathrm{avg}}(x')}. \quad (3)$$

Note here that L_g^{eff} enters as the range over which the integral to obtain τ_T is performed (in the integral, $x' = x - x_1$, where x_1 corresponds to the starting point of the effective gate). In figure 10(a), we indicate the physical beginning and ending positions of the 10 and 50 nm gates. It is apparent that the electrons start accelerating well before they reach the position where the gate actually starts. This is understandable by observing that they are actually responding to the electric field generated by the gate, and that field can extend considerably beyond the physical gate region in the regime that we are exploring.

An important question is that of the best way to determine L_g^{eff}. Wu *et al* [8] suggested that the length of the depletion region be used as L_g^{eff}, and then combined this length with an average velocity to compute a cutoff frequency. Here we take the opposite approach. Given the cutoff frequency from the small-signal RF analysis and the average velocities from the Monte Carlo simulations, we take L_g^{eff} to be the length that satisfies equation (3). Moreover, we use the positions where the electrons start to be accelerated as the starting point of the effective gates. The dots superimposed onto the velocity traces in figure 10(a) are the data points over which the integrations were performed. Note that equation (3) becomes satisfied when *the position of maximum velocity is reached*. This seems to correlate with the position at which significant numbers of carriers are transferred to the satellite valleys (recall figure 5(b)). This correlation between the end of the effective gate and the position of peak velocity appears to hold for all the cases that we have considered.

The effective gate lengths for the 10 and 50 nm cases are also indicated in the figure. Importantly, while longer than the physical gate lengths, these lengths are considerably shorter than the estimated depletion lengths, which are indicated in figure 10(b). As is evident, these extend well beyond the velocity peak. From our analysis, it is clear that using the effective gate length derived from the depletion length would lead one to *underestimate the actual cutoff frequency by a considerable amount*: a factor of 2 or more. Why is there such

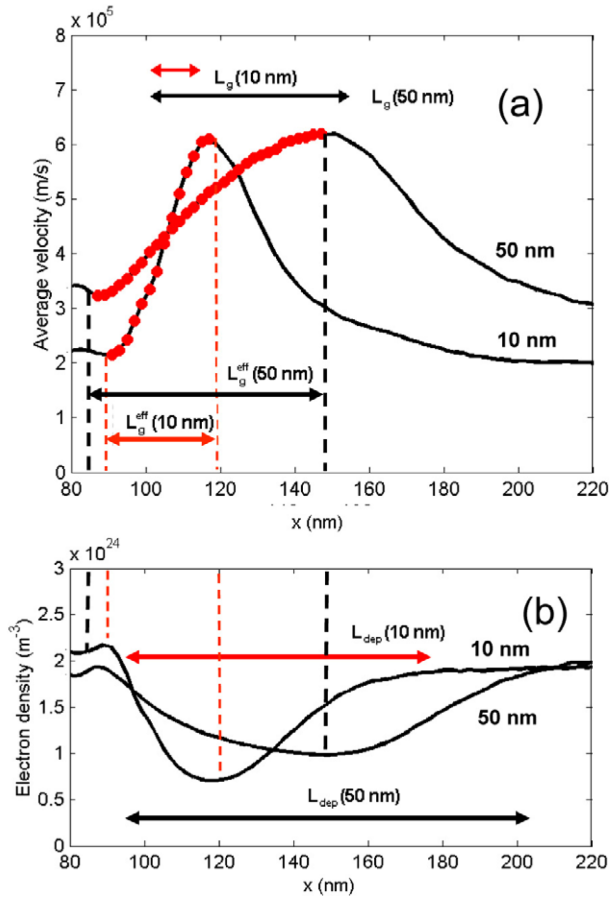

Figure 10. (a) Average electron velocity along the channel with the actual and effective gate lengths also indicated (the gray and black dashed lines indicate the beginning and end points for the 10 and 50 nm gate length cases respectively). The dots represent the data points over which the integral for transit time was computed for each, as per equation (2). (b) The corresponding average electron density with estimated depletion lengths indicated for the same two cases.

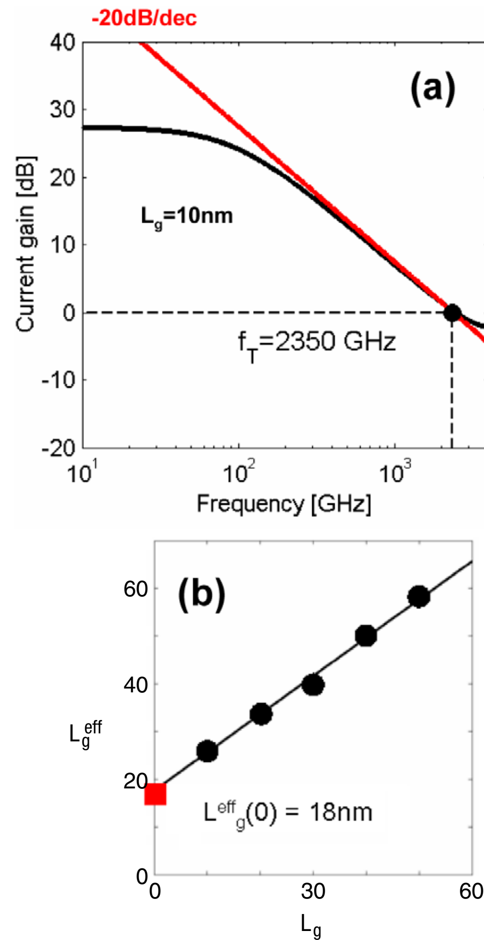

Figure 11. (a) Frequency response characteristics for a 300 nm device with $L_g = 10$ and $d_c = 10$ nm. (b) Effective gate length versus physical gate length for the set of devices with $d_c = 10$ nm. The intersection point $L_g^{\mathrm{eff}}(0) = 18$ nm gives the lower limit for effective gate length.

a large discrepancy? Importantly, if the gate is lengthened, one would expect the effective gate length as we have defined it and the depletion length to approach close to the same value when the region over which the velocity drops off becomes small compared to the physical gate length. In view of this, using the latter to estimate f_T is acceptable if the gate is *comparatively long*. It is specifically in the short gate limit that we are focusing on here where it becomes problematic.

In figure 11(a), the resulting frequency response for a 10 nm gate length and a narrower channel thickness, $d_c = 10$ nm, is shown. The linear fit using a -20 dB/decade slope in this case yields an $f_T = 2.4$ THz, a bit larger than that for the $d_c = 18$ nm case, but the difference is not profound. Indeed, as is evident in figure 12, the f_T values for the two channel thicknesses track each other rather closely.

Figure 11(b) shows the values of L_g^{eff} that we computed for the $d_c = 10$ nm case as a function of the physical gate length, L_g obtained using the methodology described above. Note that, to a very good approximation, L_g^{eff} scales linearly with L_g. Given this, one can extrapolate down to $L_g = 0$. As shown, the linear fit crosses zero at $L_g^{\mathrm{eff}} = 18$ nm. A similar analysis carried out for the $d_c = 18$ nm case yields $L_g^{\mathrm{eff}}(0) = 17$ nm [6]. We believe that the slightly lower value is due to the fringing fields being more significant in the case of the narrower channel. Given these numbers, one can make an estimate of the upper limit of f_T in either case, as we shall now show.

Figure 12 shows a plot of f_T versus $1/L_g$ for $d_c = 18$ and 10 nm. As one might expect from the previous discussion on why an effective gate length is needed, f_T does not scale linearly with the physical gate length in either case, and a far more linear trend is observed if $1/L_g^{\mathrm{eff}}$ is plotted. The intersection of the fit with the line corresponding to $L_g^{\mathrm{eff}}(0) = 17$ nm yields an estimate of the absolute upper limit for $f_{T,}$ of $\sim$2.9 THz for $d_c = 18$ nm. Meanwhile, the extrapolated upper limit for f_T is $\sim$3.1 THz using the point of intersection with the $L_g^{\mathrm{eff}}(0) = 18$ nm line. It is higher, but not hugely so, in part because of the offsetting effect of the increase in $L_g^{\mathrm{eff}}(0)$.

5. Conclusions

We have conducted systematic investigations of ultra-submicron-gate, $In_{0.52}Al_{0.48}As/In_{0.75}Ga_{0.25}As/In_{0.52}Al_{0.48}As/$

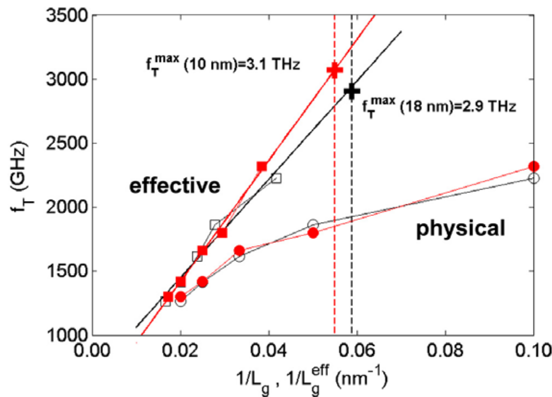

Figure 12. Cutoff frequency determined from the frequency response of the scaled devices as a function of the gate length. The black/open symbols are for an 18 nm channel and the red/solid symbols are for a 10 nm channel. Circles are for the physical gate lengths, squares are for the effective gate length. The thin solid lines indicate the linear fits to the effective gate data. The dashed lines indicate the positions $1/L_g^{\mathrm{eff}}(0)$ for the two cases, and the intersection points give the upper limit that f_T can reach for the two cases.

InP, delta-doped, pseudomorphic HEMTs using a full-band cellular Monte Carlo simulator. Our simulation work on this important class of microwave transistors has suggested high frequency device performance well above 1.5 THz for gate lengths of 20 nm and less with ~3 THz or higher perhaps being achievable according to our limit studies. Device contacts and source–drain spacing play a critical role in limiting the static DC behavior and small-signal RF response, with the effect of an *effective* internal source resistance via simulated contact doping and that of internal series resistance via source–drain spacing being a key issue. Reducing overall device length improves performance. Our simulations also show that the cutoff frequency also increases significantly with a reduction in gate length, but an important detail as regards those calculations is that the gate to channel distance also needs to be scaled (recall that we used $d_g = L_g/5$) at the same time in order to get the most dramatic enhancement. In some preliminary work that we did, we found that the results were far less impressive if d_g was not being scaled concurrently with L_g. Thus, an important aspect of this work lies primarily in the realization that scaling of this particular class of devices must be conducted in a very careful and deliberate manner. One cannot, for example, arbitrarily reduce the gate length in the hope of dramatically improving the frequency response. We found that reducing the channel thickness does not have a dramatic effect on the cutoff frequency; however our work does indicate that improved response may be achieved by reducing the source to gate distance. Given such considerations, we believe that there is certainly room for improvement over the limits that we established in the previous section. Many other aspects of the devices that could yield further improvement have yet to be studied.

In closing, we should also point out that second-order effects, such as gate to channel tunneling, do not appear in the full-band CMC simulation package that we use. When we reach the very small spacings of 2 nm that accompany the 10 nm gate length in the gate length scaling study, this is probably an effect that needs to be considered, and may impact the performance of the devices at very high frequencies. These effects will be explored in subsequent work.

Acknowledgment

This work was supported in part by the DARPA SWIFT project through the Army Research Laboratory under cooperative agreement W911NF-06-2-0012.

References

[1] Schwierz F and Liou J J 2003 *Modern Microwave Transistors: Theory, Design and Performance* (New Jersey: Wiley)
[2] Mei X B, Yoshida W, Deal W R, Liu P H, Lee J, Uyeda J, Dang L, Wang J, Liu W, Li D, Barsky M, Kim Y M, Lange M, Chin T P, Radisic V, Gaier T, Fung A, Samoska L and Lai R 2007 *IEEE Electron Device Lett.* **28** 470
[3] Yamashita Y, Endoh A, Shinohara K, Hikosaka K, Matsui T, Hiyamizu S and Mimura T 2002 *IEEE Electron Device Lett.* **23** 573
[4] Pötz W and Vogl P 1981 *Phys. Rev.* B **24** 2025
[5] Ayubi-Moak J S, Ferry D K, Goodnick S M, Akis R and Saraniti M 2007 *IEEE Trans. Electron Devices* **54** 2327
[6] Akis R, Ayubi-Moak J S, Faralli N, Ferry D K, Goodnick S M and Saraniti M 2008 *IEEE Electron Device Lett.* **29** 306
[7] Ferry D K, Ayubi-Moak J, Akis R, Faralli N, Saraniti M and Goodnick S M 2008 *J. Phys.: Conf. Ser.* **109** 012001
[8] Wu Y-R, Singh M and Singh J 2006 *IEEE Trans. Electron Devices* **53** 588
[9] Saraniti M and Goodnick S M 2000 *IEEE Trans. Electron Devices* **47** 1909
[10] Hackbush W 1985 *Multigrid Methods and Applications* (Berlin: Springer)
[11] Wigger S J, Saraniti M and Goodnick S M 1999 *MSM99: Proc. 2nd Int. Conf. on Modeling and Simulation of Microsystems (Puerto Rico, PR, April 1999)* pp 380–83
[12] Chelikowsky J R and Cohen M L 1976 *Phys. Rev.* B **14** 556
Chelikowsky J R and Cohen M L 1984 *Phys. Rev.* B **30** 4828
[13] Lee M J G and Falicov L M 1968 *Proc. R. Soc.* A **304** 319
[14] Pandey K C and Phillips J C 1974 *Phys. Rev.* B **9** 1552
[15] Kopf Ch, Kosina H and Selberherr S 1997 *Solid.-State Electron.* **41** 1139
[16] Mikkelsen J C and Joyce J B 1983 *Phys. Rev. Lett.* **49** 1412
[17] Groenen J, Carles R, Landa G, Guerret-Piecourt C, Fontaine C and Gendry M 1998 *Phys. Rev.* B **58** 10542
[18] Kunc K and Nielsen O H 1979 *Comput. Phys. Commun.* **17** 413
[19] Price D L, Rowe J M and Nicklow R M 1971 *Phys. Rev.* B **3** 1268
[20] Borcherds P H and Kunc K 1978 *J. Phys. C: Solid State Phys.* **11** 4145
[21] Ridley B K 1977 *J. Phys. C: Solid State Phys.* **10** 1589
[22] Akis R, Shifren L, Ferry D K and Vasileska D 2001 *Phys. Status Solidi* b **226** 1
[23] Jacoboni C and Lugli P 1989 *The Monte Carlo Method for Semiconductor Device Simulation* (Berlin: Springer)
[24] Grann E D, Tsen K T and Ferry D K 1996 *Phys. Rev.* B **53** 9847
[25] Hauser J R 1967 *Solid-State Electron.* **10** 577