ELSEVIER

Mathematics and Computers in Simulation 62 (2003) 357–366

MATHEMATICS AND COMPUTERS IN SIMULATION

www.elsevier.com/locate/matcom

Nonequilibrium and ballistic transport, and backscattering in decanano HEMTs: a Monte Carlo simulation study

K. Kalna*, A. Asenov

Device Modelling Group, Department of Electronics and Electrical Engineering, University of Glasgow, Rankine Building, Oakfield Avenue, Glasgow G12 8LT, Scotland, UK

Abstract

High electron mobility transistors (HEMTs) based on III–V semiconductor materials have been investigated as these devices are scaled down to gate lengths of 120, 90, 70, 50 and 30 nm. A standard Monte Carlo (MC) method coupled with the solution of Poisson's equation is employed to simulate a particle transport. The average particle velocity and the field–momentum relaxation time are studied in detail along the pseudomorphic HEMT (PHEMT) channel for two possible approaches to scaling. Nonequilibrium and ballistic transport dominate at gate lengths of 120 and 70 nm. However, velocity saturation is observed in the 50 nm gate length PHEMT which is due to strong scattering including backscattering. In addition, single and double delta doping designs are also compared. Our work indicates that the 70 nm double doped PHEMT is the most suitable design to further increase the device transconductance.

PACS: 85.30.De; 85.30.Tv; 72.15.Lh; 72.20.Ht

Keywords: Monte Carlo device simulations; High electron mobility transistors; Average velocity; Performance; Ballistic transport; Backscattering

1. Introduction

High electron mobility transistors (HEMTs) can be further scaled down into decanano dimensions in an effort to attain better performance in RF applications. However, as dimensions of the HEMTs are reduced, nonequilibrium and, particularly, ballistic transport starts to play an important role. This paper investigates electron transport in a set of scaled pseudomorphic HEMTs (PHEMTs) with a low indium content channel. The simulations have been carried out with our Monte Carlo (MC) device simulator which uses finite elements to solve Poisson's equation in the device and an enhanced electron MC transport model to simulate the carrier dynamics.

* Corresponding author. Tel.: +44-141-330-4792; fax: +44-141-330-4907.
E-mail address: kalna@elec.gla.ac.uk (K. Kalna).
URL: http://www.elec.gla.ac.uk/kalna/

0378-4754/02/$ – see front matter © 2002 IMACS. Published by Elsevier Science B.V. All rights reserved.
PII: S0378-4754(02)00223-9

The basic principles of MC method used to simulate the carrier transport are briefly repeated in Section 2. In Section 3 the salient features of our MC device simulator are outlined. The study itself is explained in Section 4 where transport characteristics for two scaling approaches are presented. Conclusions are left to Section 5.

2. Monte Carlo method for transport simulation in semiconductors

MC methods [1] are widely used in various fields of physics including nuclear physics, solid-state physics and statistical physics in general. When applied to carrier transport in solids, MC provides exact numerical solution of the Boltzmann transport equation (BTE) without necessity to solve BTE directly. MC method does not really solve BTE but the obtained distribution function $f(\boldsymbol{k}, t)$ is identical with the distribution function satisfying BTE [2,3]. This method is one of the most popular simulation technique, because it enables us to obtain exact numerical solutions of BTE using relatively simple and very effective program algorithms when compared with direct numerical techniques. Simultaneously, MC provides satisfactory microscopic interpretation of simulated processes. The essence of MC method is to simulate motion of an ensemble of carriers in $\boldsymbol{k}$ space as well as in $\boldsymbol{r}$ space. The motion of each carrier is governed by semiclassical equations of motion and by stochastic collisions with various perturbations (phonons, ions, other carriers). These collisions cause instantaneous transitions between unperturbed Bloch states with transition probabilities given by Fermi's golden rule. Using these probabilities the carrier free-flight time, the scattering channel, and the final states after scattering can be generated by random numbers. During the simulation any physical quantity which is dependent on $\boldsymbol{k}$ and $\boldsymbol{r}$ can be calculated.

Historically, the stationary carrier transport in homogeneous bulk semiconductors was first investigated by a single-particle MC simulation [4]. Over the several years nonparabolicity effects [5], diffusion [6], high-field transport with Pauli exclusion principle [7], and transient and inhomogeneous transport [8] were comprehensively included in the MC formalism.

More complicated effects such as the space charge, carrier–carrier interactions, and recombination can be studied by a many-particle (ensemble) MC method [9–11]. Introduction of carrier–carrier scattering into the ensemble MC method is formidable task. Simulations including two-particle carrier–carrier scattering due to the screened Coulomb interaction [12,13] as well as carrier–plasmon interaction were developed to examine many-body carrier dynamics [12]. Thus, ensemble MC enables us to simulate single-particle scattering processes (carrier–phonon, carrier–plasmon and carrier–ion scattering) together with two-body collision processes (carrier–carrier scattering). Another way to include inter-carrier Coulomb interactions is the molecular dynamics technique [14]. Although this technique is fully classical, it involves both short- and long-range Coulomb interactions and makes no assumptions on the screening (RPA, static screening, etc.). The technique has to be coupled with the MC simulation of "single-particle scattering" processes [14].

The starting point of the MC program is the definition of the physical system of interest including the parameters of the material and the values of physical quantities such as a lattice temperature T and an electric field $\boldsymbol{F}$. The choice of the dispersion relation $E(\boldsymbol{k})$ usually depends on the simulated transport problem. For weak fields the parabolic dispersion law is used, whereas the nonparabolic law $\hbar k^2/(2m) = E(1 + \alpha E)$ has to be used for high electric fields [4,10]. For extremely high electric fields it is necessary to calculate the band structure by pseudopotential methods [15]. At this level we also define parameters that control the simulation such as the duration of each subhistory, the desired precision of a result, etc. The next step in the program is the preliminary calculation of each scattering rate as a function of the electron energy. This step will provide information on the maximum value of these functions which will be useful for optimising the efficiency of the simulation.

Steady-state transport can be simulated by the single-particle simulation but the simulation time must be long enough so that a sufficiently representative number of particle states is sampled. The choice of simulation duration is a compromise between the need for ergodicity ($t \to \infty$) and the need for efficient use of computer time. The longer the simulation time, the less influence the initial conditions will have on average results. However, in order to avoid the undesirable effects of an inappropriate initial choice and to obtain better convergence, an elimination of the first part of the simulation from statistics may be advantageous. When the simulation aims for a study of the transient phenomenon and/or transport processes in an inhomogeneous system (e.g. electron transport in a very small device), which is exactly our case, then it is necessary to simulate many electrons. In this case the distribution of the initial electron states for the particular physical situation under investigation must be taken into account and the initial transient becomes an essential part of a result.

The subsequent step is the generation of time of free flight. The electron wave vector, $\boldsymbol{k}$, changes continuously during the free flight due to an applied electric field $\boldsymbol{F}$. Thus, if $\lambda[\boldsymbol{k}(t)]$ is the scattering probability for an electron in the state $\boldsymbol{k}$ during the small time interval $\mathrm{d}t$ then the probability that the electron, which already suffered a scattering event at time $t = 0$, has not yet suffered further scattering after time t is

$$\exp\left[-\int_0^t \mathrm{d}t'\lambda[\boldsymbol{k}(t')]\right] \tag{1}$$

which, generally, gives the probability that the interval $(0, t)$ does not contain a scattering event. Consequently, the probability $P(t)$ that the electron will suffer its next scattering event during $\mathrm{d}t$ is

$$P(t)\mathrm{d}t = \lambda[\boldsymbol{k}(t)]\exp\left[-\int_0^t \mathrm{d}t'\lambda[\boldsymbol{k}(t')]\right]\mathrm{d}t. \tag{2}$$

The free-flight time t can be generated from the equation

$$r = \int_0^t \mathrm{d}t' P(t'), \tag{3}$$

where r is a random number between 0 and 1. Once the electron free flight has terminated a scattering mechanism has to be selected. The weight of the ith scattering mechanism (when n scattering mechanisms are present) is given by

$$P_i(\boldsymbol{k}) = \frac{\lambda_i(\boldsymbol{k})}{\lambda(\boldsymbol{k})}, \qquad \lambda(\boldsymbol{k}) = \sum_{i=1}^{n}\lambda_i(\boldsymbol{k}). \tag{4}$$

After generating the random number r between 0 and 1 and testing inequalities

$$\sum_{i=1}^{j-1}\frac{\lambda_i(\boldsymbol{k})}{\lambda(\boldsymbol{k})} < r < \sum_{i=1}^{j}\frac{\lambda_i(\boldsymbol{k})}{\lambda(\boldsymbol{k})}, \quad j = 1, \ldots, n \tag{5}$$

we accept the ith mechanism if the jth inequality is fulfilled. It should be noted that the discussed selection of the free-flight time and the scattering channel can be substantially simplified by introducing a self-scattering λ_0 [16].

Once the scattering mechanism that caused the end of the electron free flight has been determined, a new state, $\boldsymbol{k}_{\mathrm{f}}$, must be chosen as the final state of the scattered electron. If the free flight ended with the self-scattering, $\boldsymbol{k}_{\mathrm{f}}$ must be taken as equal to $\boldsymbol{k}_i$, the state before scattering. When, in contrast, the true scattering has occurred then $\boldsymbol{k}_{\mathrm{f}}$ must be generated stochastically according to the differential cross section of that particular mechanism.

The last step of a simulation is the collection of statistical averages. In the ensemble MC simulation of the steady-state transport the time average is performed as follows. If N is the whole number of simulated particles we may obtain the ensemble average value of the quantity $Q(\boldsymbol{k})$ (e.g. the drift velocity, the mean energy, etc.) during the single history of duration t as

$$\langle Q(\boldsymbol{k})\rangle = \frac{1}{t}\int_0^t \mathrm{d}t' \frac{1}{N}\sum_{j=1}^{N} Q[\boldsymbol{k}_j(t')] = \frac{1}{t}\frac{1}{N}\sum_i \int_0^{t_i} \mathrm{d}t' \sum_{j=1}^{N} Q[\boldsymbol{k}_j(t')], \tag{6}$$

where j is the particle index. The integral in Eq. (6) over the time t has been separated into the sum of integrals over all free flights of duration t_i. When the steady state is investigated, t should be taken sufficiently long so that $\langle Q\rangle$ in (6) represents an unbiased estimator of average of the quantity Q over the electron gas.

In the ensemble MC simulation of transient effects it is also possible to compute the instantaneous mean value $\langle Q(t)\rangle$ as

$$\langle Q(t)\rangle = \frac{1}{N}\sum_{j=1}^{N} Q[\boldsymbol{k}_j(t)], \tag{7}$$

where $Q(\boldsymbol{k})$ can be the electron energy $E(\boldsymbol{k})$, group velocity $\boldsymbol{v}(\boldsymbol{k})$, etc.

3. Monte Carlo device simulations

Our MC device simulator uses quadrilateral finite elements [17] to represent the complex geometry of PHEMTs. A highly adaptive mesh allows us to accurately calculate electrostatic effects caused by the gate and recess geometries. The MC module includes electron scattering with polar optical phonons; inter- and intra-valley non-polar optical phonons; acoustic phonons and ionized and neutral impurity scattering. Alloy scattering and strain effects [18] are also taken into account in the InGaAs channel.

Following [19], all scattering rates and the generation of final states are modified with the form factor G (overlap integral) given by

$$G(E_{\mathrm{i}}, E_{\mathrm{f}}) = \frac{(1+\alpha_{\mathrm{i}}E_{\mathrm{i}})(1+\alpha_{\mathrm{f}}E_{\mathrm{f}}) + 1/3\alpha_{\mathrm{i}}E_{\mathrm{i}}\alpha_{\mathrm{f}}E_{\mathrm{f}}}{(1+2\alpha_{\mathrm{i}}E_{\mathrm{i}})(1+2\alpha_{\mathrm{f}}E_{\mathrm{f}})}, \tag{8}$$

where an electron with initial energy E_{i} leaves with final energy E_{f} after scattering and where α_{i} and α_{f} are the non-parabolicity parameters for the electron in its initial and final valleys, respectively. Note that nonparabolic dispersion is used to represent the band structure and to calculate the scattering rates.

The transport model has been verified by simulating drift velocities and average energies in bulk materials [20]. The MC device simulator itself has been painstakingly calibrated against a real 120 nm gate length PHEMT designed and fabricated within our department. The calculated I_{D}–V_{D} characteristics agree very well with experimental data as shown in [20]. The MC simulator may be then used to study

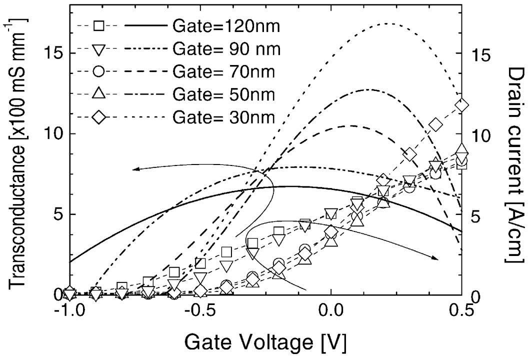

Fig. 1. I_D–V_G characteristics and transconductance for the fully scaled PHEMTs.

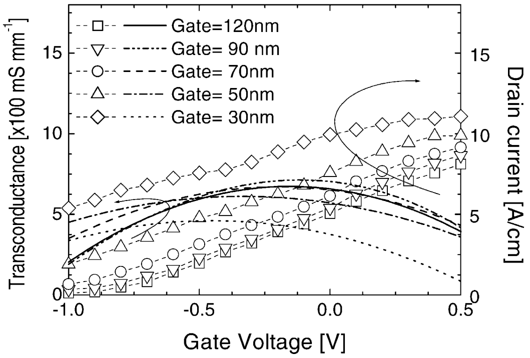

Fig. 2. I_D–V_G characteristics and transconductance for the lateral-only scaling PHEMTs.

different approaches to PHEMT scaling. The first is a full scaling when the device is scaled in both lateral and vertical directions in respect to gate lengths of 90, 70, 50 and 30 nm and the second is a device scaling in the lateral directions only while the vertical directions are kept to be the same. The comparison of the device transconductance from Fig. 1 with the transconductance from Fig. 2 tells us that only the fully scaled PHEMTs exhibit a dramatic improvement in the performance although external parasitics exert limitations [20]. In addition, we also investigate fully scaled PHEMTs in which a second delta doping layer has been introduced into the device structure [21]. Placement of the second delta doping below the channel improves the device linearity whereas placing the second delta doping above the original delta layer, near to the gate, can further improve the transconductance [22].

4. Nonequilibrium and ballistic transport

The detailed study of nonequilibrium transport in scaled PHEMTs requires monitoring the velocity of each carrier through the device during the MC simulation thus enabling the determination of the average particle velocity. Electron transport in the channel beneath the gate has a highly nonequilibrium character [23]. The average particle velocity achieves its peak value here and is much larger than its bulk saturation

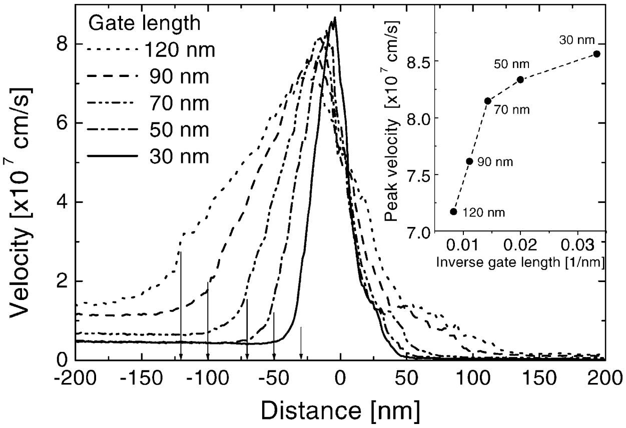

Fig. 3. Average particle velocity along the channel of fully scaled PHEMTs at $V_G = 0.0$ V and $V_D = 1.5$ V. The inset depicts the peak velocity vs. inverse gate length.

velocity. A sharp drop in the velocity is observed when electrons reach the extremely high field at the recess region on the drain side of the device (see Figs. 3 and 4).

We compared the average particle velocity along the InGaAs channel for fully (Fig. 3) and laterally (Fig. 4) scaled PHEMTs at the same applied gate and drain biases of 0.0 and 1.5 V, respectively. These biases correspond to the device in the saturation region. The average particle velocity rapidly increases when the PHEMT is fully scaled from 120 to 70 nm. However, Fig. 3 shows that the velocity saturates with the further scaling of the devices to gate lengths of 50 and 30 nm. Nevertheless, the lateral-only scaling of PHEMTs does not deliver much improvement in the average particle velocity as shown in Fig. 4. Thus it seems that lateral-only scaling is not able to deliver any increase in device performance and therefore only the full scaling really benefits device performance.

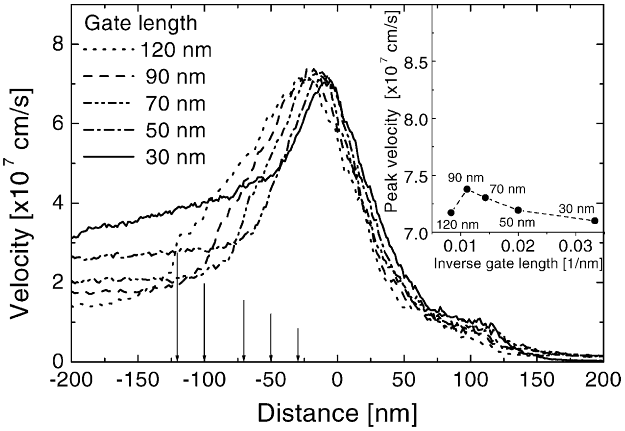

Fig. 4. Average particle velocity along the channel of lateral-only scaling PHEMTs at $V_G = 0.0$ V and $V_D = 1.5$ V. The inset shows the peak velocity vs. inverse gate length of the devices.

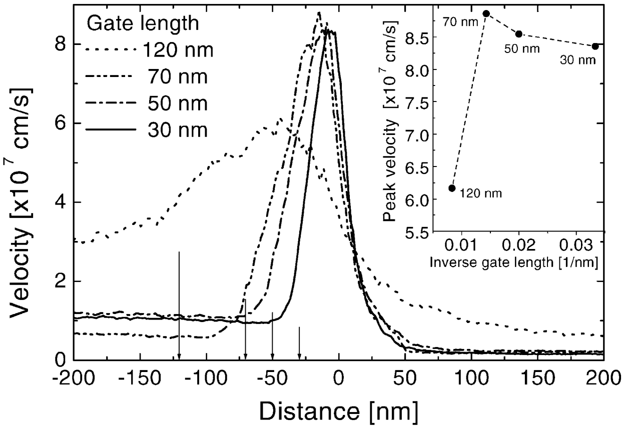

Fig. 5. Average particle velocity along the channel of double doped PHEMTs when the second delta doped layer is below the channel at $V_G = 0.0$ V and $V_D = 1.5$ V. The peak velocity vs. inverse gate length is extracted in the inset.

Comparison of the average particle velocity in the single and double doped structures is not so straightforward even we compare both devices in the saturation region. The double doped PHEMTs at $V_G = 0.0$ V and $V_D = 1.5$ V have lower pinch off and hence their 'electrostatic' stage is not the same as in the single doped device. The average velocity through the channel of the 120 nm double doped PHEMT in Figs. 5 and 6 is lower than in the 120 nm single doped PHEMT. However, in 70 nm devices the velocity becomes larger indicating that these PHEMTs are the most suitable candidates for the second delta doping layer design. There is an increase of only 4% in the peak velocity in those devices with an additional delta doping below the channel (see Fig. 5) and this rises to 16% in the devices with the second doping layer above the original doping (see Fig. 6). Both double doped devices keep their larger channel velocity at the 50 nm gate length but at the 30 nm gate length the single doped fully scaled PHEMT is slightly faster. These observations suggest that the second delta doping layer placed below the channel, which increases

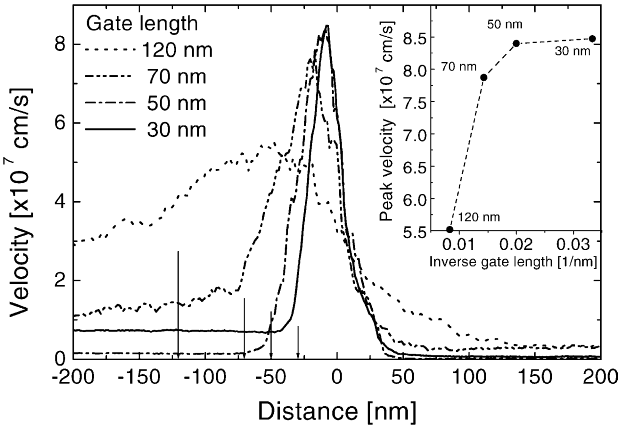

Fig. 6. Average particle velocity along the channel of double doped PHEMTs when the second delta doped layer is above the original doping near to the gate at $V_G = 0.0$ V and $V_D = 1.5$ V. The inset shows again the peak velocity.

the carrier sheet density in the device by about 70%, does not much improve the electron transport in the channel. These devices also exhibit a larger device linearity but show no improvement in transconductance. The double doped PHEMT with the additional delta doping close to the gate exhibits an increase in the transconductance compared to the single doped structure. This is consistent with the larger average channel velocity in this device.

The device gate length of decanano PHEMTs becomes comparable to the inelastic mean-free path of carriers. Hence, electrons travelling through the gate region should have a high probability of passing through this region ballistically. To study ballistic transport in the scaled devices we monitor particles in the gate-controlled-channel region [23] and then calculate the field–momentum (F–m) relaxation time as the reciprocal of λ_{Fm} given by

$$\lambda_{Fm} = \frac{e}{\hbar}\frac{|\boldsymbol{F}|}{|\boldsymbol{k}|}, \tag{9}$$

where $\boldsymbol{F}$ is the electric field vector at the particle position and $\boldsymbol{k}$ is the particle wavevector. This relaxation time represents the time during which the absolute particle momentum is relaxed due to the effect of the electric field at the particle position. The mean F–m relaxation rate is found by averaging the F–m relaxation rate, λ_{Fm}, over the number of particles passing through the gate-channel region and over some time of interest.

The mean F–m relaxation time can be compared among different devices in order to assess the typical transport situation in a selected region of a device. When the F–m relaxation time increases a large number of carriers can travel ballistically due to high electric fields and the small amount of scattering. On the other hand, a decrease in this relaxation time clearly indicates that carriers undergo many scattering events in the selected device region even if high electric fields are present. Using the field–momentum relaxation time as one of the device characteristic parameters can, with the help of all the other information acquired by MC simulation, expose the ballistic limit [24] which is anticipated as a result of the scaling process.

Fig. 7 shows the mean F–m relaxation time as a function of the inverse gate length for both scaling approaches applied to PHEMTs and GaAs HEMTs. Because only the fully scaled PHEMTs exhibit an

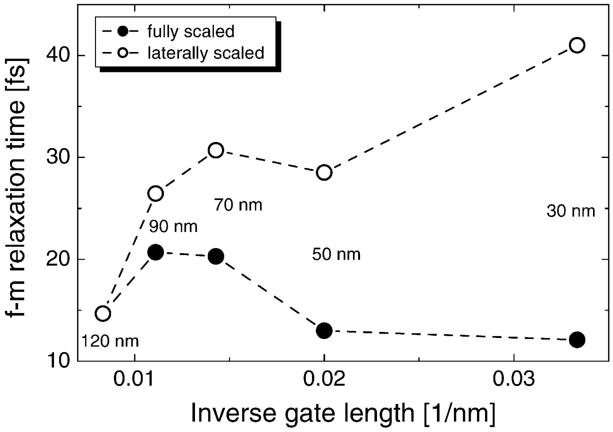

Fig. 7. Field–momentum relaxation time as a function of the inverse gate length in the single doped PHEMTs for the two scaling approaches.

improvement in transconductance we will focus on them first. The F–m relaxation time maximum is observed at the 90 nm gate length device in Fig. 7 in accordance with the behaviour of the average velocities in the channel and then the relaxation time decreases at 50 nm, finally saturating at 30 nm. This can be explained as follows. The gate-fringing effect plays a significant role in particle kinetics [25]. The impact of this effect on particles increases when the gate length is scaled down. The huge electric fields in the recess region surround the gate and, consequently, the particles are accelerated by these fringing fields when leaving the gate region on the drain side. The acceleration by the gate-fringing effect [25], however, is limited in devices with gate lengths less than 90 nm. This limitation is imposed at high energies by the increased scattering with phonons which may result in backscattering [24]. Consequently, the mean field–momentum relaxation time starts to drop rapidly and then saturates when the gate length is scaled from 50 to 30 nm. The saturation of the F–m relaxation time occurs as the field particle acceleration and the energy losses due to the increased intervalley scattering and backscattering become balanced. Both intervalley scattering and backscattering adversely affect device performance and neutralise the benefits of ballistic transport.

The behaviour of the F–m relaxation time for the lateral-only scaling PHEMTs differs from those which are fully scaled. Here, the F–m relaxation time consistently increases with increasing mean value of the electron wave vector in the channel due to less scattering. Nevertheless, although the ballisticity of the transport improves during the lateral-only scaling, device performance deteriorates because the gate has already lost control over the carriers in the channel.

Fig. 8 compares the F–m relaxation times of the single doped PHEMTs with the double doped structures for both placements of the additional delta doping layer. The double doped PHEMTs clearly suffer much more scattering up to the 70 nm gate length. But from 50 nm, transport in the double doped PHEMTs with the second doping layer placed above the original doping becomes more ballistic that transport in the single doped PHEMTs. Nevertheless, the F–m relaxation time for the double doped structures with the second doping layer placed below the channel remains greater than that for the single doped ones indicating reduced scattering.

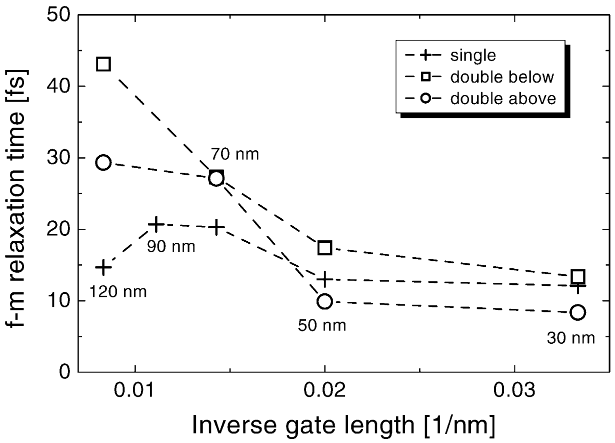

Fig. 8. Field–momentum relaxation time as a function of the inverse gate length for single and both double doped PHEMT structures.

5. Conclusions

We have carried out self-consistent device simulations of PHEMTs solving the non-linear Poisson equation and simulating particle transport for the solved potential via the MC method. The performance of PHEMTs has been investigated when these devices are scaled down in respect to gate lengths of 120, 90, 70, 50 and 30 nm. Two approaches for device scaling have been studied: the full scaling in both lateral and vertical directions and lateral-only scaling. The monitored average velocities and field–momentum relaxation times have revealed that only the full scaling of PHEMTs can improve their performance. However, the desired ballistic transport hits a ballistic limit at the 50 nm gate length even for these fully scaled devices. In addition, double doped fully scaled PHEMTs with two possible placements of the second delta doping layer have been also examined. Their average velocities and field–momentum relaxation times have answered the question why the 70 nm gate length PHEMT is the most suitable device for the placement of the additional delta doping layer.

Acknowledgements

Authors of this work kindly acknowledge the support from EPSRC under Grant No. GR/M93383.

References

[1] N. Metropolis, S.M. Ulam, J. Am. Stat. Assoc. 44 (1949) 335;
N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, J. Chem. Phys. 21 (1953) 1087.
[2] A. Reklaitis, Phys. Lett. 88A (1982) 367.
[3] D.K. Ferry, Semiconductor Transport, Taylor & Francis, New York, 2000, p. 225.
[4] W. Fawcett, D.A. Boardman, S. Swain, J. Phys. Chem. Solids 31 (1970) 1963.
[5] W. Fawcett, C. Hilsum, H.D. Rees, Solid State Commun. 7 (1969) 1257.
[6] W. Fawcett, in: A. Salam (Ed.), Electrons in Crystalline Solids, IAEA, Vienna, 1973, p. 531.
[7] S. Bosi, C. Jacoboni, J. Phys. C 9 (1976) 315.
[8] J.G. Ruch, IEEE Trans. Electron. Devices ED-19 (1972) 652.
[9] P.A. Lebwohl, P.J. Price, Solid State Commun. 9 (1971) 1221.
[10] C. Jacoboni, L. Reggiani, Rev. Mod. Phys. 55 (1983) 645.
[11] M.V. Fischetti, S.E. Laux, Phys. Rev. B 38 (1988) 9721.
[12] P. Lugli, D.K. Ferry, Physica B 117 (1983) 251.
[13] M. Moško, A. Mošková, Phys. Rev. B 44 (1991) 10794.
[14] D.K. Ferry, A.M. Kriman, M.J. Kann, R.P. Joshi, Comput. Phys. Commun. 67 (1991) 119.
[15] M.L. Cohen, J.R. Chelikowsky, Phys. Rev. B 14 (1976) 556.
[16] H.D. Rees, Phys. Lett. A 26 (1968) 416.
[17] S. Babiker, A. Asenov, J.R. Barker, S.P. Beaumont, Solid-State Electron. 39 (1996) 629.
[18] Ch. Köpf, H. Kosina, S. Selberherr, Solid-State Electron. 41 (1997) 1139.
[19] D. Matz, Phys. Rev. 168 (1968) 843.
[20] K. Kalna, S. Roy, A. Asenov, K. Elgaid, I. Thayne, Solid-State Electron. 46 (2002) 631.
[21] K.Y. Hur, K.T. Hetzler, R.A. McTaggart, D.W. Vye, P.J. Lemonias, W.E. Hoke, Electron. Lett. 32 (1996) 1516.
[22] K. Kalna, A. Asenov, in: H. Ryssel, G. Wachutka, H. Grünbacher (Eds.), Proceedings of ESSDERC 2001, Frontier Group, Nürnberg, 2001, p. 380.
[23] K. Kalna, A. Asenov, Semicond. Sci. Technol. 17 (2002) 579.
[24] M. Lundstrom, Z. Ren, S. Dutta, in: J. Faricelli, P. Leon (Eds.), Proceedings of SISPAD'00, Seattle, USA, 2000, p. 1.
[25] J. Han, D.K. Ferry, Solid-State Electron. 43 (1999) 335.