Full-band CMC simulations of terahertz HEMTs

D. K. Ferry, J. Ayubi-Moak, R. Akis, N. Faralli, M. Saraniti, and S. M. Goodnick
Department of Electrical Engineering and Arizona Institute for Nano-Electronics
Arizona State University, Tempe, AZ 85287-5706 USA
E-mail: ferry@asu.edu

Abstract. High-electron mobility transistors (HEMTs) have become an important device for high frequency and low noise applications. The performance of these devices has been pushed into the range of several hundred GHz for \( f_T \). One question that has been asked is just how high a frequency can be obtained with these devices. To study this question, 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. We have concentrated on pseudomorphic HEMTs with the structure (from the substrate) InP/InAlAs/InGaAs/InAlAs/InGaAs, with the quantum well composed of In_{0.75}Ga_{0.25}As, and have studied gate lengths over the range 10-70 nm. Various source-drain spacings have also been studied, and the performance of scaled devices 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.

Introduction
In the search for sources and detectors in the terahertz regime, the high-electron mobility transistor (HEMT) has become one of the key semiconductor devices [1]. Recently, the use of In-rich InGaAs on InP-based p-HEMTs has resulted in submillimeter-wave amplifiers above 300 GHz [2] and individual transistor cutoff frequencies approaching 600 GHz [3]. Previous simulations indicate 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 [4]. More recently, we showed that 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 p-HEMT [5]. Through an appropriate definition of an effective gate length and extrapolation, we find that this is 2.8 THz. This is far higher than some previous estimates of this quantity. In that regard, using the results of a full RF analysis [1] is crucial in formulating a definition of the effective gate length. Using the depletion length as the effective gate length [6], and using that to estimate \( f_T \) can lead to an underestimate of the theoretical value of this quantity by more than a factor of two, as our more mathematical analysis has shown.

In this paper, we will review the current understanding of the physics that is important in ultrashort gate length HEMTs, using the In-rich InGaAs channel HEMT as our prototypical device. Here, we will give some new results on scaling the gate length, discuss the role of series resistances, and the importance of the effective gate length. The latter allows the estimation of the limit from the

---

1 To whom any correspondence should be addressed.
extension to the case of a vanishing physical gate length. The importance of the contact resistance, source-drain spacing, gate-channel spacing, and quantum well width will all be discussed.

The prototypical device structure

Our particle-based simulator uses a full-band Cellular Monte Carlo approach [7]. A detailed description of how it is used to simulate a HEMT with a strained-InGaAs quantum well is given in [4], and the device structure is illustrated in Fig. 1. Other than the changes in the device dimensions, the same parameters for scattering and band structure, as well as the methodology is used here. The device structure used in our 2D simulations is the same as in ref. 4. The thicknesses and doping concentrations of the heterolayers were chosen to closely match those of fabricated devices [2]. 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 cap layer In0.53Ga0.47As/In0.52Al0.48As and channel In0.75Ga0.25As/In0.52Al0.48As heterojunctions, with 0.34 eV used at the In0.52Al0.48As/InP interface. A $\delta$-doping layer concentration of $3.5 \times 10^{12}$ cm$^{-2}$ was spread over 2 grid cells along the positive $y$-direction (normal to the channel), 1 nm from the channel. The system is represented on a 150 × 148 tensor product grid with uniform grid spacings ($\Delta x = 2$ nm) in the direction of the conduction path, while non-uniform $\Delta y$ spacing, varying between 0.5 and 14 nm, is used in the perpendicular direction. Finer gridding ($\Delta y = 0.5$ nm) is used across the active channel region and the highly-doped cap layers, while a coarser mesh is used in the buffer and substrate regions. The starting position of the gate for each device was nominally fixed at just over 100 nm from the source, although a variation in this will be mentioned below. In scaling studies, the gate to channel distance was scaled, $d_g = L_g/5$. Finally, the quantum well channel is nominally 18 nm thick, although we will discuss briefly a 10 nm channel and its effect on the scaled devices.

Figure 1. Schematic layout of the device being simulated.

The small-signal RF analysis is performed by first biasing each device at a point in the operating region of the $I_d-V_d$ characteristics and allowing it to reach a quasi-static dc (steady-state) condition. These bias points were centered about the position of peak transconductance. Next, a 200 mV step is applied directly to the gate contact while holding the drain at a constant electrostatic potential. The corresponding drain currents are recorded after each update of the field equations and
used during post-simulation analysis to compute the high-frequency response (via Fourier transform), which is then used to extract the cutoff frequency.

**General results**

One of the most important aspects is the role of contact resistance on device performance. In our simulator, we control this property by adjusting the doping concentration in the small region which is defined to be the contact. It is found that using too low a value of the doping leads to potential drops in the contact regions, and subsequent poorer device performance. We established previously that the actual simulated contact resistance can be related to observed experimental values \([4]\), so that good agreement with experimental \(I_d-V_d\) characteristics can be obtained. More importantly, 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. In the left-hand panel of Fig. 2, we map the position of the carriers, located between the source and the gate, in the full Brillouin zone. Here, it may be observed that these carriers are almost entirely low energy carriers in the \(\Gamma\) valley of the conduction band. Hence, these carriers are moving relatively slowly and lead to the effective series resistance. This 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 Fig. 3, 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 nm to 60 nm produces almost a 40% increase in the transconductance. This change in distance also affects the gate-source capacitance, so that the cutoff frequency 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.

![Figure 2](image-url)

**Figure 2.** The projection of the particles onto the Brillouin zone. The left-panel depicts those particles located between the source and the gate. The right-panel depicts those particles located between the gate and the drain. The color is an energy code (which is measured relative to the valence band maximum).

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 \(\Gamma\) valley is so low, the density of states is also low and the resulting scattering rate from L (or X) to \(\Gamma\) is quite small \([8]\). We see this in the right-hand panel of Fig. 2, where we plot the position of the carriers, located in the channel between the gate and the drain, in the Brillouin zone. Some carriers remain in the \(\Gamma\) 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 percent 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 valleys (something like 1.5 the L valley to valence band maximum gap), so that there is very little impact ionization. These devices can take high bias.

4. Scaling the gate length
The scaling of these HEMTs is studied by looking at a set of 5 devices, whose gate lengths vary from 10 to 70 nm. In each case, the source-drain spacing is held to 0.3 \( \mu \)m, but the gate-to-channel separation is scaled with the gate length according to \( L_{gc} = L_{G}/5 \). In the left panel of Fig. 4, we plot the velocity along the channel as a function of position within the channel. It should be noted that the velocity begins to rise before the gate is reached, a result of field spreading in the gate direction [9]. The peak of the velocity, in each case, occurs before the end of the gate is reached, and this is the point

**Figure 3.** Comparison of the transconductance of a 10 nm gate length, 10 nm channel device with a total source-drain spacing of 0.3 \( \mu \)m, as the source-gate length is varied.

**Figure 4.** The left panel illustrates the velocity along the channel for a set of scaled devices (discussed in the text), while the right panel illustrates the transconductance in these same devices for \( V_{DS} = 1.0 \) V.
at which significant numbers of carriers are being transferred to the satellite valleys, as mentioned above. In the right panel of Fig. 4, 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 infers a better frequency response as well.

The small-signal RF analysis was performed by biasing each device at a point in the operating region of the $I_d-V_d$ characteristics and allowing it to reach a steady-state condition. These bias points were centered about the position of peak transconductance. Next, a 200 mV step was applied directly to the gate contact while holding the drain at a constant electrostatic potential. The corresponding drain currents are recorded after each update of the field equations and used during post-simulation analysis to compute the corresponding small-signal gains. The resulting values for the cutoff frequencies $f_T$ are shown in Fig. 5 as a function of the physical gate length. This cutoff frequency corresponds to a particular value of the transit-time, or delay time, in the important part of the transistor. This latter time is related to the velocity in the channel through [5]

$$f_T = \frac{1}{2\pi f_T} = \int_0^{L_g} \frac{dx'}{v_{avg}(x')} \cdot (1)$$

As discussed above, the carriers begin to accelerate prior to arriving at the gate. Given the cutoff frequency from the small-signal RF analysis and the average velocities from the Monte-Carlo simulations, we seek a length that satisfies (1). By performing this integration over the discretized mesh used in the simulation, we arrive at the remarkable result that (1) becomes satisfied when the position of maximum velocity is reached. This means that we can define an effective gate length as the distance from the initial acceleration point to the position of maximum velocity. Moreover, by plotting this effective gate length as a function of the physical gate length, we can extrapolate to a zero of the latter, and reach the ultimate limiting frequency of the transistors. In Fig. 5, we also show the performance ($f_T$) as a function of the effective gate length and indicate the limiting frequency obtained from this analysis.

Figure 5. Cutoff frequency determined from the frequency response of the scaled devices as a function of the gate length. The open symbols are for a 18 nm channel and the closed symbols are for a 10 nm channel. The solid line is for the physical gate length. The dashed line and extrapolation limit are for the effective gate length determined according to (1).
Conclusions
We have studied pseudomorphic HEMTs, based upon an In-rich InGaAs channel, and growth upon an InP substrate. We have found that outstanding performance can be achieved in these devices. To achieve good performance, it is necessary to reduce the source-gate distance as well as the source-channel distance. Study of a set of scaled devices, in which the gate-to-channel distance is scaled with the gate length has shown performance limits above a terahertz. Nevertheless, the determination of an effective gate length in these devices illustrates that the fringing field on the source side of the gate is important and that further improvement could be obtained by further reduction in the source-gate distance. On the other hand, one may not want to reduce the gate-drain separation equivalently, as the present devices achieve some protection against impact ionization through this drift length in which the majority of the carriers are in the satellite valleys where they have much higher ionization thresholds.

Another interesting point that arises from Fig. 4 is that the thickness of the quantum well in these devices is not very important. As discussed, we studied devices with thicknesses of 10 nm and 18 nm. The former had slightly higher transconductances, but also had larger values of the effective gate length at the shortest gate lengths. While the 10 nm channel devices extrapolated to a higher ultimate cutoff frequency (3.1 THz vs. 2.9 THz), this resulted more from the more linear variation of the effective gate length with physical gate length. In fact, a glance at the axes shows that the extrapolated effective gate length in the 10 nm channel is larger than in the 18 nm channel. This probably means that the thinner channel results in a larger extension of the fringing field toward the source. This points out the importance of trying to reduce the source-gate spacing to limit this fringing field. In fact, use of a self-aligned gate (and source) would yield considerable high frequency improvement. Thus, it would be a mistake to interpret the “ultimate” limits in Fig. 4 as being true upper limits. A more properly scaled device, with reduced source-gate, and perhaps source-drain, spacing would most likely lead to an increase in the limits shown in Fig. 4.

We should also point out that second-order effects, such as gate-to-channel tunnelling, 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, this is probably an effect that needs to be considered, and may impact the performance of the devices at very high frequencies.

Acknowledgments
This work was supported in part by the DARPA SWIFT project through the Army Research Laboratory under cooperative agreement W911NF-06-2-0012. The authors wish to acknowledge many helpful discussions with R. Lai and B. Boos.

1. F. Schwierz and J. J. Liou, Modern Microwave Transistors: Theory, Design and Performance, (John Wiley & Sons, Inc., New Jersey, 2003).
2. X. B. Mei et al., IEEE Electron Dev. Lett. 28, 470 (2007).
3. Y. Yamashita et al., IEEE Electron Dev. Lett. 23, 573 (2002).
4. J. S. Ayubi-Moak et al., IEEE Trans. Electron Dev. 54, 2327 (2007).
5. R. Akis et al., submitted for publication.
6. Y.-R. Wu, M. Singh, and J. Singh, IEEE Trans. Electron Dev. 53, 588 (2006).
7. M. Saraniti and S. M. Goodnick, IEEE Trans. Electron Dev. 47, 1909 (2000).
8. E. D. Grann, K. T. Tsen, and D. K. Ferry, Phys. Rev. B 53, 9847 (1996).
9. J. R. Hauser, Sol.-State Electron. 10, 577 (1967).