Physics and Mitigation of Excess OFF-State Current in InGaAs Quantum-Well MOSFETs

Jianqiang Lin, Student Member, IEEE, Dimitri A. Antoniadis, Life Fellow, IEEE, and Jesús A. del Alamo, Fellow, IEEE

Abstract—A number of recent reports have noted excess OFF-state leakage current ($I_{OFF}$) in scaled InGaAs quantum-well nMOSFETs. There is growing evidence that a combination of hand-to-hand tunneling (BTBT) and a floating-body bipolar gain effect is responsible for this. Unless this issue is effectively addressed, the scaling potential of this transistor structure will be compromised. This paper presents a detailed study of the physics of $I_{OFF}$ and explores $I_{OFF}$ reduction strategies through 2-D device simulations that have been calibrated with experiments. In essence, under OFF conditions at even moderate values of $V_{ds}$, a BTBT process at the drain-end generates holes in the channel and thereby reduces the source–channel potential barrier. This results in injection of electrons into the channel that contribute to enhanced $I_{OFF}$ while the holes are injected into the source where they recombine. In a nanoscale device, the bipolar effect that is at play here can have a very large current gain and amplify many fold even a small BTBT current. A study of approaches to mitigating this effect is analyzed here. It is concluded that the most effective strategy is to minimize the bipolar current gain rather than BTBT in scaled transistors.

Index Terms—III–V, band-to-band tunneling (BTBT), bipolar effect, floating body, MOSFETs, quantum-well (QW), self-aligned.

I. INTRODUCTION

InAs-RICH InGaAs is a promising channel material for future CMOS applications due to its superior electron transport properties [1], [2]. In recently demonstrated scaled InGaAs quantum-well (QW) MOSFETs with tight source and drain (SD) spacing around the gate, excess drain-to-source leakage current has been observed that prevents the effective turn-OFF of the transistor [3]–[6]. This is a serious problem that must be addressed before the logic potential of this material system can be realized. As a reference, the International Technology Roadmap for Semiconductors (ITRS) requires $I_{OFF}$ of 10 pA/μm for low-standby-power applications [7].

Recently, there has been mounting evidence that excess OFF-state current has its origin in band-to-band tunneling (BTBT) at the drain-end of the channel [3]–[6]. In [3], $I_{OFF}$ was found to feature the characteristic voltage and temperature signatures of gate-induced drain leakage. It was postulated that $I_{OFF}$ arises from BTBT at the drain-end of the channel [3]. A more detailed study has recently shown that while BTBT indeed is the likely triggering mechanism, a parasitic bipolar transistor action in the floating QW-channel is responsible for significant amplification of the BTBT current [6]. The simulations in [6] indicate that the bipolar current gain in nanoscale devices can be rather large, easily above 100, and it rises rapidly with channel length reduction. This greatly magnifies the BTBT current and makes the excess $I_{OFF}$ problem much worse.

It has been well known for some time that in floating-body silicon-on-insulator (SOI) MOSFETs, the presence of a lateral bipolar junction transistor (BJT) with substantial current gain yields a dramatic enhancement of BTBT and impact-ionization current [8], [9]. A number of anomalies have been traced to this combination of mechanisms, such as the kink effect for analog device applications [10], jitter [11], and premature breakdown [12], among others [10]. In InGaAs high-electron-mobility transistors, floating body effects associated with hole generation through impact ionization have also been widely reported [13]–[15]. These have been found to result in the kink effect, excess gate leakage, excess and frequency-dependent output conductance, and premature breakdown and burnout [16].

Multiple factors can influence the magnitude of BTBT and bipolar gain in InGaAs QW-MOSFETs, and a systematic study of this is yet to be performed. The purpose of this paper is to fill this gap through a simulation study that has been calibrated with experiments. As compared with [6] whose focus was the experimental verification of the coupling of BTBT and the bipolar gain effect, this paper targets detailed physical understanding and exploration of mitigation strategies.

This paper is organized as follows. Section II summarizes the physics of BTBT and bipolar amplification in QW-MOSFETs. In Sections III and IV, we present the simulations of the OFF-state current that illuminate BTBT-induced source–channel barrier lowering and the bipolar gain amplification and its $L_{g}$ dependence. Section V describes a study of device design approaches to mitigate the OFF-state current problem. In particular, we study the role of: 1) doping of the channel region; 2) InAs composition of channel and cap; 3) carrier lifetime; and 4) geometrical design of the access regions. Finally, Section VI provides the conclusions of this paper.
This paper should be instrumental in identifying effective off-state current reduction strategies for future nanoscale InGaAs QW MOSFETs.

II. PHYSICS OF EXCESS $I_{\text{OFF}}$ IN SCALED InGaAs QW-MOSFETS

Experimental evidence of bipolar amplification of BTBT in floating-body QW-MOSFETs was recently presented in [6]. The basic concept is summarized in Fig. 1, which shows a schematic cross section of a simplified InGaAs QW-MOSFET biased in the off-state with high $V_{ds}$ and a corresponding energy band diagram; in the rest of this paper, we will refer to this bias condition simply as the off-state. Overlaid are arrows that indicate electron and hole flow.

In the off-state, BTBT takes place in the high-field region around the drain-end of the channel. As sketched in Fig. 1 by the thin arrows, valence electrons from the channel tunnel across this region and are collected by the drain. The holes that are left behind accumulate in the channel leading to the reduction of the source–channel energy barrier, which under steady-state conditions allows the BTBT-generated holes to be injected into the source and recombine in the heavily doped cap or at the source–contact interface.

An important consequence of the reduction of the source–channel energy barrier is electron injection from the source into the channel. These electrons diffuse through the channel and contribute an additional drain current component. This is indicated in Fig. 1 by the thick blue arrow. In essence, this is a bipolar gain effect similar to what has been observed in floating-body SOI MOSFETs [8], [9]. In a nanoscale device, the gain of the parasitic bipolar transistor can be quite large and, as a result, the total drain current can be many times higher than the BTBT electron–hole generation current.

III. SIMULATION SETUP

A cross-sectional schematic of the simplified InGaAs QW-MOSFET modeled in this paper is shown in Fig. 2. This is a self-aligned device with a raised SD architecture similar to devices in which excess drain current has been observed in the off-state [3]–[6]. In our model device, the gate dielectric is characterized by equivalent oxide thickness (EOT) $= 1 \text{ nm}$. The channel is an 8-nm thick $\text{In}_x\text{Ga}_{1-x}\text{As}$ with InAs composition of $x_{\text{ch}}$ on an InAlAs buffer layer lattice matched to InP. The cap that forms the raised SD is also $\text{In}_x\text{Ga}_{1-x}\text{As}$ but with an InAs composition of $x_{\text{cap}}$. The doping level is $2 \times 10^{19} \text{ cm}^{-3}$, and no doping or $\delta$-doping is present elsewhere. The thickness of the cap and channel is denoted as $t_{\text{cap}}$ and $t_{\text{ch}}$, respectively. The space between the gate edge and the ohmic contacts is known as the access region and has a length $L_{\text{access}}$. Our baseline structure uses $x_{\text{ch}} = 0.7$, $x_{\text{cap}} = 0.53$, $t_{\text{cap}} = 30 \text{ nm}$, $t_{\text{ch}} = 8 \text{ nm}$, $L_{\text{access}} = 0$, and $L_g = 40 \text{ nm}$.

The 2-D device simulator Sentaurus Device by Synopsys [17] was used in these simulations. The coupled Poisson, electron and hole continuity equations are solved self-consistently with a nonlocal-path BTBT model. Quantum effects are not included in these simulations due to convergence difficulties. We expect some minor corrections when quantum effects are accounted for as a result of the presence of a quantum dark region that pushes the charge centroid away from the front and back interfaces of the channel. In the baseline simulation setup, only recombination at the SD metal contacts, which are characterized by an infinite surface recombination velocity, is allowed. The contacts are long enough for their length not to be relevant. Recombination in the body of the channel and the cap is neglected because the carrier diffusion lengths in these two regions are estimated to be much longer than the gate length and the cap thickness, respectively. This estimation is based on the experimental carrier lifetimes of In$_{0.53}$Ga$_{0.47}$As/InP structures [18] and reasonable mobilities [19], [20], and yields a diffusion length that is over $10 \mu\text{m}$ in the channel and $\sim 0.2 \mu\text{m}$ in the $n^+$ cap. Nevertheless, we explore the effect of carrier recombination in the body of the semiconductor in Section V.

For this paper, we apply the TCAD models and parameters in [6] as they match the experimental results reasonably well.

IV. OFF-STATE CURRENT MODELING

This section gives the results of simulations of devices with different $L_g$ values, where BTBT is turned on and off.
The goal is to provide physical understanding of the relevant physics.

A. BTBT-Induced Source–Channel Barrier Lowering

Fig. 3(a) shows the simulated subthreshold $I_d$–$V_{gs}$ characteristics at $V_{ds} = 0.6$ V for $L_g = 40$, 100, and 800 nm MOSFETs with and without BTBT. For all devices, $I_d$ decreases monotonically as $V_{gs}$ is reduced when the BTBT model is not included, as expected in an ideal MOSFET. The subthreshold swing degradation with decreasing $L_g$ is due to short-channel effects (SCEs). In contrast, when the BTBT model is included, the off-state drain current, $I_d$ ($V_{gs} = −0.2$ V, $V_{ds} = 0.6$ V), reaches a minimum value and then increases as $V_{gs}$ decreases far below threshold. This level of $I_d$ is set by the coupling between BTBT, bipolar gain, and gate voltage through the modulation of the height of the potential barrier, $\phi_b$, that exists between the source and the channel under the gate (Fig. 1).

To better understand this, Fig. 4 plots the conduction band edge ($E_c$) from source to drain in the longitudinal direction at a depth of 2 nm into the channel ($y = −2$ nm in Fig. 2) for two transistors with $L_g = 800$ and 40 nm. With the BTBT model turned off (Fig. 4, left panels), as $V_{gs}$ decreases, $q\phi_b$ increases until normal hole accumulation in the channel becomes prevalent. In contrast, with the BTBT model present (right panels), $q\phi_b$ increases only initially with decreasing $V_{gs}$ and then becomes independent of $V_{gs}$. This is what causes $I_d$ in the off-state to essentially saturate. Making $V_{gs}$ more negative flattens the top of the inverse-U-shaped band profile in the channel. This is because a significant concentration of holes accumulates in the channel leading to a more effective screening of the longitudinal charge distribution.

The evolution of $q\phi_b$ with $V_{gs}$ for the simulations of Fig. 3(a) is more clearly visible in Fig. 3(b). In the absence of BTBT, $q\phi_b$ increases as $V_{gs}$ becomes more negative. Eventually, saturation takes place when the Fermi level at the surface approaches the valence band edge and the channel goes into hole accumulation. Including BTBT introduces increased accumulation of holes that limits the rise of $q\phi_b$ at higher values of $V_{gs}$ as follows. Reduction in $V_{gs}$ gives rise to higher $V_{dg}$ and hence higher electric field at the drain edge of the channel [3]; this enhances the BTBT generated hole current injected into the channel ($I_{h,BTBT}$) which in turn brings about a reduction in $q\phi_b$ counterbalancing the effect of $V_{gs}$.

B. Bipolar Gain Amplification and $L_g$ Dependence

As shown in Fig. 3(b), the source–channel barrier height in the off-state in the presence of BTBT is essentially the same for the $L_g = 100$ and 800 nm devices, which is understandable because the BTBT current ($I_{BTBT} = |I_{e,BTBT}| + |I_{h,BTBT}|$) only depends on the E-field at the drain edge and is independent of gate length. This is indeed verified in Fig. 5 that plots $I_{BTBT}$ corresponding to the three gate lengths. Compared with the difference between total currents $I_{T}$, the $I_{BTBT}$ currents are very similar—almost the same for the two longer devices and within a factor of 3, including the shortest one.

The bipolar current gain, $\beta$, on the other hand, does depend on $L_g$. $\beta$ can be defined as

$$\beta = \frac{I_e}{I_{BTBT}}.$$
In the subthreshold regime, electron transport through the channel takes place by diffusion and a shorter $L_g$ leads to a steeper electron gradient and higher electron current. Fig. 5(a) also plots the electron channel current ($I_e$) and the terminal drain current ($I_{d,T}$). $I_{d,T}$ denotes the total current by source electron injection and BTBT, $I_{d,T} = I_e + I_{BTBT}$. It is clear that in spite of a similar $I_{BTBT}$, the shorter devices have significantly higher $I_e$ and, therefore, $I_{d,T}$. For short devices, since $\beta$ is much larger than unity ($I_e \gg I_{BTBT}$), it is expected that $I_{d,T} \simeq I_e$. The gain reduces for longer devices. This is indeed the case for the longer device, $L_g = 800$ nm, as shown in Fig. 5(b).

Our simple physical arguments lead us to postulate an inverse linear relationship between $\beta$ and $L_g$. Indeed, this is what we observe in the simulations at all values of $V_{gs}$ in Fig. 6, which also shows that, for a given $L_g$, $\beta$ is reduced as the gate voltage becomes more negative. The physical origin is discussed as follows.

SCEs complicate this idealized picture in a number of ways. In a BJT, the quasi-neutral base is the portion of the base with nearly zero longitudinal E-field, through which injected carriers flow by diffusion. Due to the presence of the depletion regions associated with the p-n junctions, the effective electrical thickness of the quasi-neutral base is smaller than that of the metallurgical base. In analogy with this, in a QW-MOSFET in the subthreshold regime, the portion of the channel that is under negligible lateral field is shorter than the gate length. This does not appreciably affect long-channel devices, but it becomes significant in short-channel transistors. As Fig. 4 shows, for $L_g = 800$ nm, the conduction band under the gate is rather flat, while for $L_g = 40$ nm, it exhibits an inverse-U shape. The effective channel length in this case is substantially shorter than $L_g$ leading to an increase in current gain.

We find in our simulations that for a given layer design, the difference between physical gate length and effective channel length is weakly dependent of $V_{gs}$. To the first order, we neglect this dependency. Then this difference can be seen in Fig. 6 where the extrapolation of $\beta^{-1}$ to $\beta^{-1} = 0$ yields $\Delta L_g \approx 70$ nm. $\Delta L_g$ is related to the penetration of the electric field lines that emanate from the SD into the channel.

This effectively shortens the effective channel length that electrons diffuse through on their way from source to drain.

Fig. 6 reveals that $\beta$ decreases as $V_{gs}$ is made more negative. As in an n-p-n bipolar transistor, $\beta$ is related to the barrier that holes must overcome to get injected into the emitter (source, in this case). The lower the barrier, the easier the injection and the lower $\beta$ is. In this regime, the reduction of $\beta$ as $V_{gs}$ becomes more negative, is associated with the reduction of $\phi_b$ for negative $V_{gs}$, as shown in Fig. 3(b).

### V. Transistor Design for Off-State Leakage Reduction

In this section, we discuss the device design approaches to reduce BTBT and bipolar gain. The degree to which such approaches might be employed to mitigate excess off-state $I_d$ in InGaAs QW-MOSFETs would depend on the considerations of the overall device performance and specifications. A detailed study of this is outside the scope of this paper.

#### A. Delta Doping and Channel Doping

Different doping profiles have been implemented in experimental InGaAs QW-MOSFETs. N-type $\delta$-doping is typically added beneath the channel (indicated by dashed line in Fig. 1) to reduce on-state access resistance and to maintain high channel mobility [3], [21]. Uniform p-type doping of the channel region has also been used [22]. We have found that the doping type and profile strongly influence the BTBT rate and the bipolar gain characteristics, as well as many other important transistor parameters, and should be an important consideration in transistor design.

Fig. 7 shows the simulated subthreshold $I_d-V_{gs}$ and $I_{BTBT}-V_{gs}$ characteristics for $L_g = 40$ nm for different (a) backside $\delta$-doping type and doping level and (b) channel uniform p-type doping level. For both simulations, $V_{ds}$ is 0.6 V.
n-type and blue with a right-pointing arrow for p-type. In Fig. 7(b), the arrows indicate channel p-doping level changes from 0 to $3 \times 10^{18}$ cm$^{-3}$.

Changing doping level and type affects the threshold voltage, which shifts positively with increasing p-type doping and negatively with increasing n-type doping. In addition, the saturated $I_d$ (in the OFF-state) increases as the n-type doping increases and it decreases as p-type doping increases though eventually this reduction saturates. Interestingly, the behavior of $I_{BTBT}$ is counter to this, it decreases as n-type doping increases, and it increases as p-type doping increases.

Fig. 7 reveals the important role that $\beta$ plays in OFF-state $I_d$ and its dependence on doping. This can be seen more clearly in Fig. 8 that plots $\beta$ at $V_{gs} = -0.4$ V and $V_{ds} = 0.6$ V against the doping level in the $\delta$-doped layer (top x-axis) and channel p-type doping $N_A$ (bottom x-axis).

The large impact of channel doping on the bipolar gain can be understood by considering the role of the vertical electric field across the channel. In the channel of a QW-MOSFET biased in the subthreshold regime, an electric field exists in the transverse direction (perpendicular to the surface). Its presence can significantly influence the bipolar action that is partially responsible for the excess OFF-state current. To illustrate this, Fig. 9 shows the energy band diagram in the transverse direction under the gate in a device biased in the OFF-state with some degree of hole accumulation in the channel. The presence of a transverse field splits the electron and hole distributions spatially with holes being located closer to the surface and the electrons placed against the channel–buffer interface. Effectively this also implies an energy separation between the electron and the hole populations that, neglecting quantization, is narrower than the bandgap by an amount roughly equal to the band bending in the channel $q\phi_c$ in Fig. 9. In analogy to the impact on $\beta$ of bandgap narrowing in the base of a BJT, this should increase the current gain by an amount roughly given by $\exp(q\phi_c/kT)$. Similar behavior in the bipolar action of a Si MOSFET has been reported [23].

In addition, in Si SOI MOSFETs, an increase in OFF-state $I_d$ is observed after irradiation [24], [25] that generates positive fixed charge in the buried oxide beneath the channel and leads to electron accumulation at the back interface.

The impact of channel doping on the bipolar gain shown in Fig. 8 can be easily understood in these terms. Fig. 10(a) shows energy band diagrams in the vertical direction in the channel under the gate for three different types of backside $\delta$-doping at the same $V_{gs} - V_t$. For clarity, the band diagrams are matched at the oxide–semiconductor interface. (b) $\beta$ versus $q\phi_c$ at fixed $V_{gs} - V_t = -0.35$ V and $V_{ds} = 0.6$ V in semilog scale indicates the exponential dependence of $\beta$ on $q\phi_c$. $q\phi_c$ is extracted in the center of the channel ($z = 0$).
direction in the subthreshold regime. Consequently, the device architectures that improve the gate control over the channel potential in the subthreshold regime, such as ultrathin-body QW-MOSFET, double-gate FinFET, trigate MOSFET and gate-all-around nanowire design will be an advantage.

B. InAs Composition in Channel and Cap

So far, we have studied a device design in which the InAs composition in the channel is $x_{ch} = 0.7$ and that of the n$^+$-cap is $x_{cap} = 0.53$. The use of a different InAs composition in the cap and channel results in a heterojunction at the cap/channel interface. This introduces an additional barrier to hole injection from the channel into the cap that greatly increases the bipolar current gain. In essence, this behaves as a heterojunction bipolar transistor [26]. Fig. 11 shows this case by sketching an energy band diagram under the contacts in the vertical direction.

We have studied the effect of varying $x_{cap}$ while keeping $x_{ch}$ constant. Fig. 12 graphs $I_d$, $I_{BTBT}$, and $\beta$ versus $x_{cap}$ at a fixed $V_{gs} = -0.35$ V and $L_g = 40$ nm for $x_{ch} = 0.7$. It is clear that OFF-state $I_d$ decreases when $x_{cap}$ increases from 0.3 to 1. This is because $\beta$ decreases in a very significant way as the hole barrier at the channel–cap interface is reduced and eventually eliminated. In contrast, the BTBT current increases as $x_{cap}$ increases.

BTBT is inherently a nonlocal process. At the drain edge of the channel, electron generation mostly happens in the cap, and hole generation in the channel [6]. Hence, the InAs composition of both the cap and the channel, through their respective bandgaps, affects the tunneling barrier. As the bandgap in the cap is reduced, the tunneling barrier for BTBT also shrinks and BTBT at the drain edge of the channel is enhanced. This is consistent with experimental observations in [5]. The increase in $I_{BTBT}$, however, is significantly weaker than the bipolar gain reduction and OFF-state leakage follows the trend of the latter. When $x_{cap}$ approaches 1, $\beta$ vanishes, and OFF-state leakage approaches $I_d$.

Fig. 13 shows the impact of $x_{ch}$ for a fixed $x_{cap} = 0.53$. OFF-state $I_d$ increases as $x_{ch}$ increases. In this case, $\beta$ and $I_{BTBT}$ are both enhanced. $\beta$ enhancement can be understood by the heterojunction effect and the BTBT current enhancement as a result of bandgap reduction in the channel.

This paper reveals that the InAs composition in channel and cap both play a very important role in determining OFF-state leakage. All other things being equal, it is clear that reducing the valence band discontinuity ($\Delta E_v = E_{v, ch} - E_{v, cap}$) at the channel–cap interface constitutes an effective way to suppress OFF-state current.

C. Carrier Lifetime

In SOI MOSFET, an effective approach to mitigate the parasitic bipolar effect is to introduce recombination centers that reduce the carrier lifetime [27]. Recombination of the BTBT-generated holes in the channel brings down their concentration and the need to inject them across the source–channel barrier.

This effect is studied by introducing a Shockley–Read–Hall-type recombination process in the cap and channel that is characterized by a carrier lifetime, $\tau$ [18]. Fig. 14 shows the impact of $\tau$ on OFF-state $I_d$, $I_{BTBT}$, and $\beta$ at $V_{gs} = -0.45$ V and
We have studied four different ways of accomplishing this: 1) reduction of the vertical electric field in the channel by doping engineering; 2) reduction of the heterojunction effect by engineering the InAs compositions in the cap and channel; 3) reduction of lifetime of excess carriers generated by BTBT; and 4) dimensional control of the extrinsic device.

We have studied four different ways of accomplishing this: 1) reduction of the vertical electric field in the channel by doping engineering; 2) reduction of the heterojunction effect by engineering the InAs compositions in the cap and channel; 3) reduction of lifetime of excess carriers generated by BTBT; and 4) dimensional control of the extrinsic device.

**REFERENCES**


Jianqiang Lin (S’08) received the B.Eng. (Hons.) and M.Eng. degrees in electrical engineering from the National University of Singapore, Singapore, in 2007 and 2009, respectively. He is currently pursuing the Ph.D. degree with the Massachusetts Institute of Technology, Cambridge, MA, USA. He was a Research Intern with the IBM T. J. Watson Research Center, Yorktown Heights, NY, USA, in 2012 and 2013, respectively.

Dimitri A. Antoniadis (M’79–SM’83–F’90–LF’14) received the B.S. degree in physics from the National University of Athens, Athens, Greece, in 1970, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA, in 1972 and 1976, respectively. He joined Massachusetts Institute of Technology, Cambridge, MA, USA, in 1978, where he is currently the Ray and Maria Stata Professor of Electrical Engineering.

Jesús A. del Alamo (S’79–M’85–SM’92–F’06) received the Telecommunications Engineer degree from the Polytechnic University of Madrid, Madrid, Spain, in 1980, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA, in 1983 and 1985, respectively. He has been with the Massachusetts Institute of Technology, Cambridge, MA, USA, since 1988, where he is currently a Donner Professor and a MacVicar Faculty Fellow.