Role of Joule Heating on Current Saturation and Transient Behavior of Graphene Transistors

Sharnali Islam, Zuanyi Li, Vincent E. Dorgan, Myung-Ho Bae, and Eric Pop, Senior Member, IEEE

Abstract—We use simulations to examine current saturation in sub-micron graphene transistors on SiO$_2$/Si. We find that self-heating is partly responsible for current saturation (lower output conductance) but degrades current densities above 1 mA/$\mu$m by up to 15%. Heating effects are reduced if the supporting insulator is thinned or, in shorter channel devices, by partial heat sinking at the contacts. The transient behavior of such devices has thermal time constants of $\sim$30–300 ns, which is dominated by the thickness of the supporting insulator and that of the device capping layers (a behavior also expected in ultrathin-body SOI transistors). The results shed important physical insight into the high-field and transient behavior of graphene transistors.

Index Terms—Current saturation, graphene field-effect transistor (FET), scaling, self-heating, thermal transient.

I. INTRODUCTION

GRAPHENE has attracted much interest for transistor applications, initially due to its high carrier mobility, i.e., $\sim$10$^4$ cm$^2$/V·s [1]. Recent work has also found drift velocity saturation at high field in graphene, at values several times higher than in silicon [2]. However, velocity saturation alone is not sufficient for current saturation because the carrier density can continue to increase with drain voltage in a zero-band-gap material, where the channel cannot be fully pinched off. Current saturation is important for low output conductance $g_o$ and amplifier gain [1], [3], and in practice, it has been partly achieved through a combination of velocity saturation and electrostatic charge control [4], [5]. At the same time, high-field transport in graphene is also influenced by self-heating [2], [6], as revealed by recent infrared and Raman thermal imaging [7]–[10].

In this letter, we examine the effect of self-heating on current saturation in sub-micron graphene-on-insulator (GOI) transistors through electro-thermal device simulations. We consider the role of the buried oxide thickness $t_{box}$ under the graphene and of the device length $L$. We also observe that practical graphene devices can be operated in transient (pulsed) mode and calculate their thermal time constants, i.e., the timescales over which the device temperature ramps up or cools down after electrical switching.

II. CURRENT SATURATION

The schematic of a typical GOI transistor is shown in Fig. 1(a). Our simulations are based on the drift–diffusion approach, calculating carrier densities, electric field, drift velocity, potential, and temperature along the channel and contacts self-consistently. The simulator was extensively tuned against experimental data [8], [10], including contact effects [11]. The metal–graphene contact resistance per unit area used here is $\rho_C = 111$ $\Omega$·$\mu$m$^2$, which is near the low end of the range for typical Pd–graphene or Au–graphene contacts [11]. The Dirac voltage of simulated devices is $V_D = 0$ V, and the background temperature is $T_0 = 293$ K. Other parameters are as in [2], including compact models of mobility and velocity saturation dependence on carrier density and temperature. Since carrier mean free paths in typical GOI transistors are in the 20 to 80 nm range [8], [10], the model is most reliable for devices greater than $\sim$0.1 $\mu$m.

We first investigate self-heating and current saturation in a device with channel length and width $L = W = 1$ $\mu$m. Fig. 1(b) shows the computed current–voltage ($I_D$–$V_{SD}$) of this device with $t_{box} = 300$ nm SiO$_2$ and vertical electric fields of 0.3, 0.6, and 1.0 MV/cm, respectively. The dashed lines represent current without self-heating ($T = T_0$), whereas the solid lines show current degradation when Joule heating is self-consistently taken into account. Thus, simulations suggest that self-heating is at least partly responsible for the current...
percentage of saturation current degradation with graphene device instability in experimental studies [2], [8].

channel length from 1 to 0.25

long-term device and dielectric reliability [12].

t vertical field. Dashed lines show less degradation and sublinear dependence on

Fig. 1(d) displays the temperature profiles corresponding to the current saturation but also the internal carrier distributions. [2]. Thus, self-heating at high field influences not only the saturation current as a function of \( t_{\text{box}} \) for the same three channel lengths and vertical field. Dashed lines show less degradation and sublinear dependence on \( t_{\text{box}} \) for sub-0.5-\( \mu \)m channel lengths due to heat sinking effect of contacts. saturation observed in recent experiments on devices of comparable size and bias [4], [5].

Fig. 1(c) shows the total carrier density at the highest voltage and current bias point in Fig. 1(b), with and without self-heating. Interestingly, because graphene is a gapless material, we find that significant self-heating during operation can alter the majority carrier concentration by thermal carrier generation [2]. Thus, self-heating at high field influences not only the current saturation but also the internal carrier distributions. Fig. 1(d) displays the temperature profiles corresponding to the maximum bias points for the three cases in Fig. 1(b). We note that sustained temperature rises \( \Delta T > 200 \) K have been linked with graphene device instability in experimental studies [2], [8].

We now study the peak temperature rise \( \Delta T \) and the percentage of saturation current degradation \( \Delta I/I_{\text{sat}} \) as we reduce \( t_{\text{box}} \) from 300 to 50 nm. For all \( t_{\text{box}} \), the peak \( \Delta T \) and \( I_{\text{sat}} \) are taken at the same \( V_{SD} = 2 \) V for vertical fields of 0.5, 1, and 2 MV/cm. Fig. 2(a) shows that the peak \( \Delta T \) of devices with channel length \( L = 1 \) \( \mu \)m proportionally scales with \( t_{\text{box}} \), as expected. However, we note that, even in the limit of \( t_{\text{box}} \rightarrow 0 \) (graphene device directly on substrate, which is similar to graphene on SiC), the temperature rise is nonzero due to the thermal resistance of the graphene–substrate interface and that of heat spreading into the substrate itself [2], [10].

Fig. 2(b) shows \( \Delta T/I_{\text{sat}} \) due to self-heating as a function of \( t_{\text{box}} \). As a simple guideline, a \~{}5\% degradation in \( I_{\text{sat}} \) corresponds to \( \Delta T \sim 170 \) K above room temperature. For current density near \~{}1 mA/\( \mu \)m, as for the top curve in Fig. 1(b) on \( t_{\text{box}} = 300 \) nm, the current degradation due to Joule heating can be \~{}10\%, and for higher current densities, the self-heating effect is proportionally larger. This can be partly compensated by reducing \( t_{\text{box}} \) and \( L \), as described here and in the subsequent discussion. We note that elevated temperatures not only decrease device performance but also have profound effects on long-term device and dielectric reliability [12].

We next explore the effect of Joule heating while scaling the channel length from 1 to 0.25 \( \mu \)m. Fig. 2(c) shows current-voltage curves computed with and without self-heating, indicating that the self-heating effect is less in shorter channel devices. Fig. 2(d) also plots \( \Delta I/I_{\text{sat}} \) due to self-heating versus \( t_{\text{box}} \) for the same channel lengths, at the same drain output conductance \( g_o = \partial I_D/\partial V_{SD} \). Less current degradation at shorter channel lengths is explained by an enhanced role of heat dissipation “laterally” to the contacts in addition to “vertically” through the oxide. This was also recently observed in experimental work on sub-0.5-\( \mu \)m graphene nanoribbons (GNRs) [13], which noted that heat dissipation into the contacts begins to play a role when the device length is \( \leq \sim 3 \) times the thermal healing length (\( L_{HT} \)). The thermal healing length is a measure of the lateral heat diffusion along the graphene, \( L_{HT} \approx 0.2 \) \( \mu \)m in graphene on 300 nm thick SiO\(_2\) and approximately half in GNRs, which have lower thermal conductivity [8], [13]. Increased heat loss to the contacts is also seen as a sublinear rise in current degradation, as shown in Fig. 2(d), for the shorter devices. Our present model numerically accounts for heat spreading into the substrate and the contacts [10]; however, this can be also treated to a good approximation analytically as in [13].

From a practical point of view, our simulations suggest that thermal effects are always significant in modern graphene devices [4], [5] at current densities > 0.5 mA/\( \mu \)m or lateral fields > 1 V/\( \mu \)m. To avoid this, devices could be built on substrates with thinner \( t_{\text{box}} \) or higher thermal conductivity (e.g., sapphire). However, some amount of self-heating can lead to better current saturation (lower \( g_o \)), but not carefully considering such effects can cause long-term device instability [12].

III. THERMAL TRANSIENT

While the preceding section focused on the effects of self-heating on DC characteristics, this section explores the transient device behavior. We perform finite-element (FE) simulations, as shown in Fig. 3(a) and (b), which show the temperature of one-half of the cross section marked by a dashed line in Fig. 1(a). Isothermal boundary conditions (\( T = T_0 \)) are applied 10 \( \mu \)m away from the device at the bottom and right edges of the Si substrate, and other boundaries are adiabatic. Thermal boundary resistance is accounted for at graphene interfaces with SiO\(_2\) [2], [10]. We used temperature-dependent values for the thermal conductivity and heat capacity of the oxide [14], although the effect was small, \< 5\%.

An input power of 0.5 mW is initially applied to the graphene channel and then turned off after 2.5 \( \mu \)s. Fig. 3(a) and (b) corresponds to temperature distributions at the end of the heating pulse in a device without and with a capping layer (assumed to be SiO\(_2\)), respectively. This can be roughly understood as a typical device in a laboratory setup, versus one that is integrated in a package. The temperature transient of the channel midpoint is shown in the inset in Fig. 3(c) for a capping layer \( t_{\text{cap}} = 200 \) nm and \( t_{\text{box}} = 250 \) nm. The thermal time constant \( \tau \) is obtained by fitting the temperature decay as \( T(t) = T_0 + T_1 e^{-t/\tau} + T_2 (1 + (t/t_0)^{-b}) \), where \( T_0 = 293 \) K is the base temperature, \( T_1 + T_2 \) is the steady-state peak temperature, and the third term is used to fit the long tail of the temperature decay due to the (small) residual heating transient of the Si substrate. Typical values for \( b \) are in the range of 0.5–2.5, and typical values for \( t_0 \) are from tens to hundreds of nanoseconds. In the “no cap” case, \( T_2 \) is less than 30\% \( T_1 \), but it becomes comparable with \( T_1 \) in cases “with cap.”

![Graphene device simulation results](image-url)
The symbols in Fig. 3(c) summarize the calculated thermal time constant of the graphene device, as the $t_{\text{box}}$ is scaled, for devices with a capping layer of 200 and 500 nm and without (“no cap”). We can understand the scaling of the thermal time constant through a simple analytic model, which includes each region as a lumped thermal resistance $R_{\text{th}}$ and thermal capacitance $C_{\text{th}}$. The thermal time constant $\tau$ is the sum of contributions ($\Sigma R_{\text{th}}C_{\text{th}}$) from the relevant regions, i.e.,

$$\tau \approx \left( f_1 \frac{C_V}{k_{\text{ox}}} t_{\text{box}}^2 + \left[ f_2 \frac{C_V}{k_{\text{cap}}} t_{\text{cap}} + \frac{C_V m}{k_{\text{m}}} m \right] (t_{\text{box}} + t_{\text{eq}}) \right)$$

where $C_V = 1.76 \text{ MJ K}^{-1} \text{ m}^{-3}$ and $C_V m = 2.88 \text{ MJ K}^{-1} \text{ m}^{-3}$ are the heat capacities of the oxide and metal gate [15], respectively; and $t_{\text{m}}$ and $k_{\text{m}}$ (40 W m$^{-1}$ K$^{-1}$ for Pd) are the thickness and the thermal conductivity of the metal gate, respectively.

The geometrical prefactors $f_1 \sim 0.6$ and $f_2 \sim 0.8$ represent the fractions of the total temperature drop in the bottom oxide and in the top capping layer, respectively. The last term $t_{\text{eq}} \approx 200$ nm accounts for the thermal equivalent of transient cooling in the Si substrate (the limit $t_{\text{box}} \rightarrow 0$), which is consistent with previous studies on bulk CMOS devices [16]. We note that the aforementioned analytic model could also be applied to other devices based on atomically thin materials such as MoS$_2$ or to ultrathin-body silicon-on-insulator (SOI) transistors.

The model of (1) is plotted with dashed lines in Fig. 3(c), which is in good agreement with our FE simulations (symbols). The FE results are realistic within 10%–20% accuracy, depending on the simulated domain size and the choice of 3-D versus 2-D simulations (the main tradeoff being CPU time); however, the main physical trends persist. These results suggest that thermal time constants follow an approximately quadratic dependence on $t_{\text{box}}$, which contributes to both the thermal resistance and the thermal capacitance of the device.

The capping layer and the metal gate do contribute to the term in (1) that is linear in $t_{\text{box}}$ but do not aid in “cooling” the device otherwise. Thus, a thicker gate or capping layer only adds “thermal ballast” and can increase the thermal time constant. Interestingly, due to its thinness, the graphene layer itself does not influence the thermal transient of the device, which is dominated by the heating of the surrounding materials. This is a unique aspect of devices based on graphene (or other 2-D monolayer materials such as MoS$_2$) versus that of older SOI technology, where a substantial thickness of the Si “body” retains nonnegligible heat capacity and thermal resistance [17, 18].

To conclude, we have found that Joule heating during operation is partly responsible for current saturation and degradation observed in graphene device experiments. Self-heating is reduced with thinner dielectrics, and for sub-0.5-μm channel lengths the contacts begin to play a role in heat sinking. The thermal time constants of GOI devices are in the range between 30–300 ns and are strongly dependent on the materials surrounding the channel. Thermal transients are much slower than electrical transients (~1–10 ps), which is consistent with previous work on SOI technology [17, 18]. This implies that graphene devices are slow to heat up or cool down after electrical switching, and for instance, pulsed operation on timescales shorter than the thermal time constant can benefit from reduced self-heating compared with DC operating modes.

References