Direct Observation of Self-Heating in III–V Gate-All-Around Nanowire MOSFETs

SangHoon Shin, *Student Member, IEEE*, Muhammad Abdul Wahab, *Member, IEEE*, Muhammad Masuduzzaman, *Member, IEEE*, Kerry Maize, Jiangjiang Gu, *Member, IEEE*, Mengwei Si, *Student Member, IEEE*, Ali Shakouri, *Member, IEEE*, Peide D. Ye, *Fellow, IEEE*, and Muhammad Ashraful Alam, *Fellow, IEEE*

***Abstract*— Gate-all-around (GAA) MOSFETs use multiple nanowires (NWs) to achieve target I_{ON}, along with excellent 3-D electrostatic control of the channel. Although the self-heating effect has been a persistent concern, the existing characterization methods, based on indirect measure of mobility and specialized test structures, do not offer adequate spatiotemporal resolution. In this paper, we develop an ultrafast high-resolution thermoreflectance (TR) imaging technique to: 1) directly observe the increase in local surface temperature of the GAA-FET with different number of NWs; 2) characterize/interpret the time constants of heating and cooling through high-resolution transient measurements; 3) identify critical paths for heat dissipation; and 4) detect *in situ* time-dependent breakdown of individual NW. Combined with the complementary approaches that probe the internal temperature of the NWs, the TR-images offer a high-resolution map of self-heating in the surround-gate devices with unprecedented precision, necessary for the validation of electrothermal models and the optimization of devices and circuits. In addition, we develop the simple compact model of the complex structure, which can explain experimental observations and can provide the internal temperature of the NWs.**

***Index Terms*— Gate-all-around (GAA), MOSFETs, nanowire (NW), reliability, self-heating, thermoreflectance (TR) measurement, variability.**

I. Introduction

MULTIGATE devices, such as FinFET, gate-all-around FETs (GAA-FETs) improve 3-D electrostatic control of the channel, but the corresponding increase in self-heating may compromise both the performance and the reliability. Although the self-heating effect (SHE) of FinFET appears significant, but tolerable [1], the same may not be true for the GAA geometry [2], [3], especially in quasi-ballistic regime where hot spots and nonclassical heat-dissipation pathways may lead to localized heating and damage to gate insulators.

Over the years, a number of self-heating characterization techniques have been developed. The electrical methods, such as four-terminal gate resistance [4], ac output conductance [5], [6] pulsed-I–V methods [7], [8], have been devised to characterize temperature in the channel (T_C). Briefly, the four-terminal gate resistance method uses a specialized test structure to monitor changes in the gate resistance (R_G) due to the self-heating. The temperature dependence of R_G is first calibrated by monitoring the change in R_G as a function of substrate (chuck) temperature, T_B. Subsequently, as the channel self-heats during the normal operation, one obtains the channel temperature by assuming that $R_G(T_B) = R_G(T_C^*)$. Note that the substrate heating is homogenous, but the self-heating is not; therefore, T_C^* may not equal T_C exactly. Other electrical characterizations, such as ac output conductance and pulsed I–V methods, rely on the difference between the electrical and the thermal time-constants to determine T_C. After all, the electrical response is almost instantaneous and power dissipation can follow high-frequency input signal; on the other hand, thermal response is slower and depends on heat dissipation pathways. In general, these electrical techniques are somewhat indirect due to complicated calibrations necessary to convert changes in electrical parameters (e.g., resistance and mobility) to the channel temperature. The most important limitation is that they rely on signals integrated over nanowires (NWs) and cannot provide spatially resolved temperature distribution necessary to analyze failure mechanisms in transistors and optimize heat dissipation in an IC.

In contrast, optical characterization methods, such as infrared (IR) thermography [1], [9-11], micro-Raman [12], and others, can map the spatial distribution on the surface temperature (T_S) [13], [14]. The spatial resolution for each method is defined by the diffraction limit ($\lambda/2$), which makes IR thermography inappropriate to characterize quasi-ballistic sub-100-nm GAA transistors. Other optical methods involve scanning, as opposed to imaging; the spatial resolution is high, but the methods are unsuitable to characterize transient heating and cooling. The optical technique to be discussed in this paper, namely, thermoreflectance (TR) imaging, offers a compromise: submicrometer wavelengths of various illumination sources (~400–800 nm) offer higher resolution compared with IR techniques, and large-area imaging enables mapping of transient heating and cooling at submicrosecond time-scales [15]–[17]. The NW-resolved temperature maps allow one to establish the dynamics of degradation of each NW

Manuscript received March 18, 2015; revised May 22, 2015; accepted June 3, 2015. Date of publication August 4, 2015; date of current version October 20, 2015. The review of this paper was arranged by Editor J. S. Suehle. *(Corresponding author: Muhammad Ashraful Alam.)*

The authors are with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: shin136@purdue.edu; mwahab@purdue.edu; mmasuduz@purdue.edu; kmaize@purdue.edu; Jiangjianggu@gmail.com; msi@purdue.edu; shakouri@purdue.edu; yep@purdue.edu; alam@purdue.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TED.2015.2444879

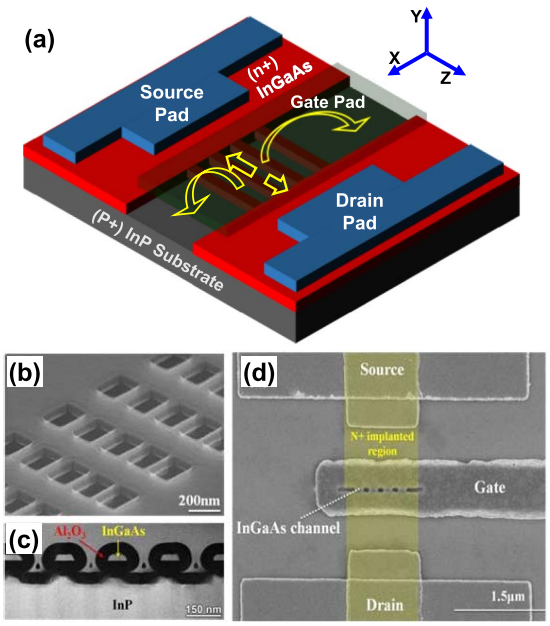

Fig. 1. (a) A schematic of an InGaAs GAA NW n-channel MOSFET. (b) SEM image of parallel NWs. (c) Side view: STEM image of the cross section of the InGaAs NWs. (d) Top view: SEM image of the parallel InGaAs NWs. The images are taken from [18].

as a function of time. Eventually, T_S must be mapped back to T_C for predictive modeling; therefore, optical techniques must be complemented by the electrical characterization within a self-consistent modeling framework for predictive modeling of the implications of self-heating on the performance and the reliability of the GAA and the FinFET transistors.

This paper is organized as follows. We first develop in Section II an ultrafast, the high-resolution TR imaging technique to directly observe the local time-dependent rise of the surface temperature, $\Delta T_S(x, y, t_d)$. In Section III, a variety of transistors with different number of NWs are explored, and the NW-dependence of self-heating is analyzed and interpreted. Indeed, this high-resolution transient measurement would allow us to characterize the time constants of heating and cooling of the channel. In Section IV, we develop a thermal compact model, which explains the experimental observations systematically and can anticipate the internal temperature (T_C) of the NWs based on the surface temperature (T_S). Finally, Section V shows how the high-spatial resolution of TR imaging offers new insights into the mechanics of correlated degradation of individual NWs as they approach dielectric breakdown (BD). We conclude this paper in Section VI by summarizing the key results.

II. Experimental Setup

A. Device Geometry

The devices used in this paper are InGaAs GAA nMOSFETs (Fig. 1), with different oxide thicknesses (T_{ox}), channel lengths (L_{ch}), and the number of NWs. The substrate is InP, and highly thermal conducting gate, source, and drain pads are used to facilitate the electrical measurement. The fabrication process is described in detail in [18] and [19], and the device dimensions are listed in Table I. The steep

TABLE I
DESCRIPTION OF THE SAMPLES ($L_{ch} = 70\sim80$ nm, AND $W_{NW} = 30$ nm) USED IN THIS PAPER [18], [19]

	Sample A IEDM 2011	Sample B IEDM 2012
Channel Material	$In_{0.53}Ga_{0.47}As$	$In_{0.65}Ga_{0.35}As$
L_{ch} (nm)	50-120	20-80
W_{NW} (nm)	30-50	20-35
H_{NW} (nm)	30	30
L_{NW} (nm)	200	200
Gate Oxide (T_{ox})	10nm Al_2O_3	3.5nm Al_2O_3
EOT (nm)	4.5	1.7
# of NW	1, 4, 9, 19	4

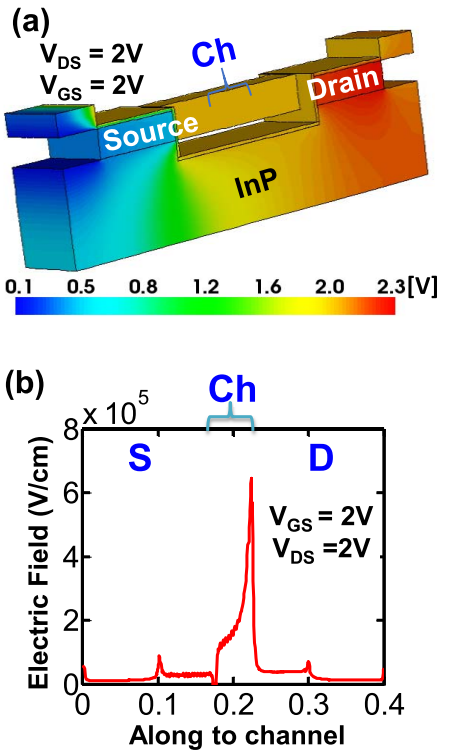

Fig. 2. (a) Simulated potential profile of the GAA MOSFET (Sample A) for $V_{GS} = 2$ V and $V_{DS} = 2$ V. Strong gate controllability over the NWs is confirmed [8]. Note that the gate (surrounding the NW) is removed for easier visualization. (b) The peak position of the electric field (in the electric field versus potential plot) confirms that the potential drop and the power dissipation are maximum at the channel-drain edge.

subthreshold slope, reported experimentally in [18] and [19], is reproduced by the 3-D Sentaurus simulation (Fig. 2), confirming excellent electrostatic control of GAA-FET [20].

B. Self-Heating and Heat Dissipation

With the application of the gate (V_{GS}) and drain (V_{DS}) biases, current flows through the channel (I_D). Heat dissipation occurs at the channel-drain edge ($P \sim I_D \times V_{DS}$). Therefore, each NW acts as a heat source and the substrate contact acts as a heat-sink [see Fig. 1(a)]. In principle, heat can also dissipate through the top surface by air convection, as well as through the gate, source, and drain metal pads. However, given

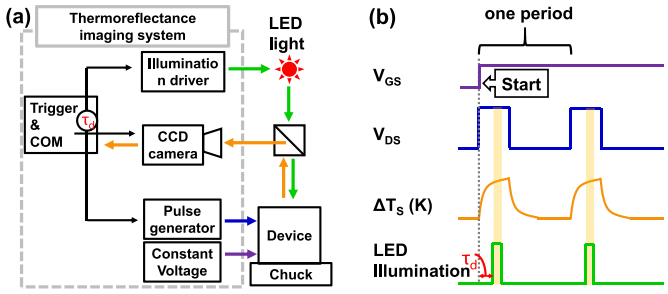

Fig. 3. (a) A schematic of TR imaging system. A pulse generator (V_{DS}) and a constant voltage source (V_{GS}) drive the transistor. A control computer triggers the illumination driver and the CCD camera for a given delay time with respect to V_{DS}. (b) Timing diagram for transient TR imaging with a given LED delay time (t_d).

the heat transfer coefficient in air is $h \sim 10$ W/m$^2/K$, the heat-flux through top surface [$F = hA\Delta T \sim 10 \times (600$ nm $\times$ 4 μm) $\times\, 100K = 2.4 \times 10^{-9}$ W] is negligible compared with heat dissipation through the substrate (10^{-3} Watts).

C. Setup and Principle of TR Measurement

During the TR imaging [17], [21], a high-speed LED pulse illuminates the device, and a synchronized chare coupled device (CCD) camera captures the reflected image [see Fig. 3(a)]. Briefly, the gate pad surface (Au) is illuminated through an LED ($\lambda = 530$ nm) via an objective lens (magnification 100). The reflected light from the surface from the gate pad is captured on a variable frame rate, 14-bit digitization, Andor CCD camera with 512×512 active pixels [15].

Theoretically, this technique relies on the change of the complex refractive index of a material with differential increase in temperature (ΔT_S), so that the change in local reflectance of the device surface is given by

$$\frac{\Delta R}{R_0} = \frac{1}{R_0} \cdot \frac{dR}{dT}\bigg|_{T=T_0} \times \Delta T_S \equiv \boldsymbol{k} \cdot \Delta T_S \qquad (1)$$

where $\boldsymbol{k}(K^{-1})$ is the TR coefficient. The calibration of $\boldsymbol{k}$ allows a CCD image to be interpreted as a map of $\Delta T_S(x, y)$, with 50-mK resolution.

For the transient measurement of $\Delta T_S(x, y, t_d)$, the device is periodically turned ON and OFF by a V_{DS} pulse train [Fig. 3(b)], allowing the channel to heat and cool, respectively. By controlling the delay of the LED pulse with respect to the beginning of the V_{DS} pulse, the TR image can capture different phases of the transient heating and cooling kinetics, with ~50-ns temporal resolution. The delay time (t_d) for the LED illumination can be varied and each illumination pulse acts as a camera shutter. Every V_{DS} cycle produces an image capturing the thermal state of the substrate at time t_d. The average of these images improves the signal-to-noise ratio and produces a high-resolution map of $\Delta T_S(x, y, t_d)$. Finally, regarding the spatial resolution, note that the sensor size is 8.2×8.2 mm^2; since the objective lens use a 100$\times$ magnification, the area imaged is 80×80 μm^2. When resolved to 512×512 pixels, each pixel corresponds to 157×157 nm^2.

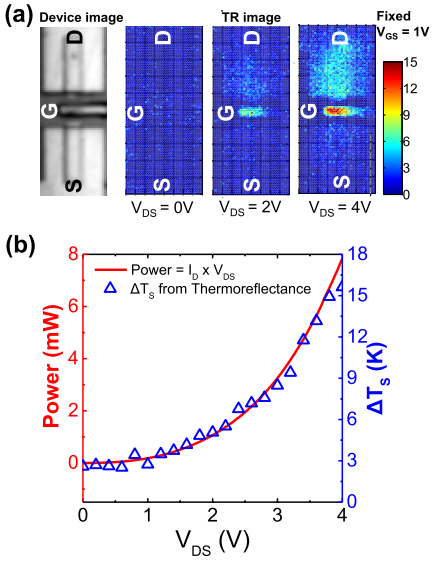

Fig. 4. (a) CCD image of the top view of the transistor and TR images under drain bias ($V_{DS} = 0$~4 V) with fixed $V_{GS} = 1$ V. (b) ΔT_S and power (heat) dissipation ($= V_{DS} \times I_D$) corresponding to (a).

D. Calibration of Thermoreflectance Coefficient, $\boldsymbol{k}$

The change in reflectivity (ΔR) of most metals under visible spectral range is proportional to the change in temperature, so that once the TR coefficient ($\boldsymbol{k}$) is obtained, ΔR can be mapped to ΔT_S [see (1)]. Unfortunately, $\boldsymbol{k}$ must be calibrated, because it depends on the wavelength, the angle of incidence, and the polarization of the incident light, as well as the surface properties of the reflecting material. The calibration is performed by heating the sample by placing it on an external microthermoelectric stage. The temperature of the sample is monitored by microthermocouple while capturing the reflection changes by the CCD camera. The $\boldsymbol{k}$ for the specific setup is obtained by plotting the change in reflectivity as a function of temperature measured by the thermocouple [22].

E. Validation of Thermoreflectance Imaging

Fig. 4(a) shows the TR images of a GAA transistor with four NWs (Sample A). Here, $V_{GS} = 1$ V, but V_{DS} changes from 0 to 4 V. The gate metal covering channel region (600×1250 nm^2) defines the calibrated Au surface that allows us to obtain the surface temperature of the device. We observe a number of features in these images. First, the images show that the asymmetry between the source and the drain temperatures increases with V_{DS}. This is expected, because power dissipated at the drain, $P \sim I_D \times V_{DS}$, increases with the drain bias. The energy is most likely dissipated close to the drain edge. The spatial extent of the temperature is related to the heat diffusion through the drain metal contact, heated by the bottom edge of the drain. Second, in the $V_{DS} = 4$ V image, there is a space between the gate and the drain, which appears unheated. This is an artifact: the segment arises from the gap between the drain and the gate metal contacts. Obviously, the channel region in the gap is heated, but the low $\boldsymbol{k}$ of the semiconductor makes the region

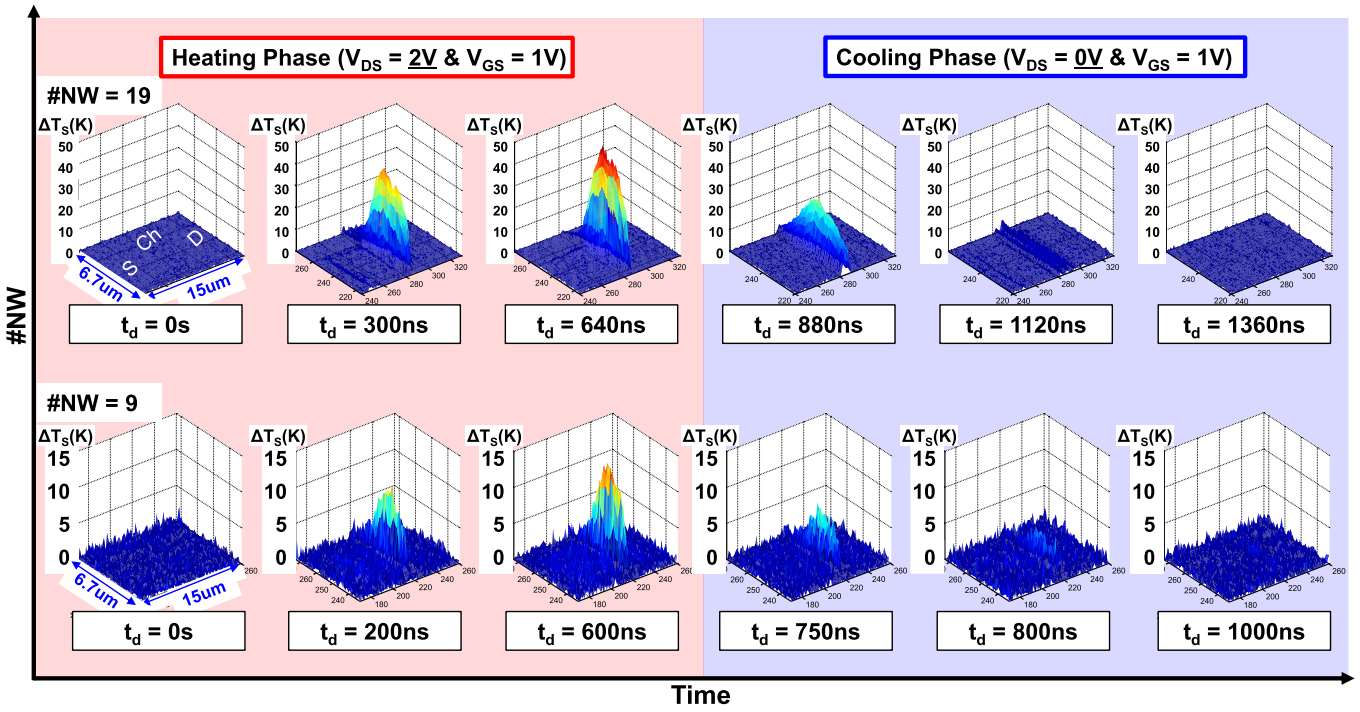

Fig. 5. 3-D TR images for heating ($V_{DS} = 2$ V and $V_{GS} = 1$ V) and cooling ($V_{DS} = 0$ V and $V_{GS} = 1$ V) phases. For clarity, only three images (out of more than 15) per cycle per device are shown. In addition, the images of four NWs are available, but not shown. The device with 19 NWs (top) shows higher saturation temperature compared with the device with 9 NWs (bottom). The heating and cooling time constants lie on the order of 100–500 ns, depending on the number of NWs, oxide thickness, and so on.

invisible in a TR map. Finally, the significant heating in the gate suggests that even for this short-channel transistor, one cannot neglect power-dissipation within the channel region. Finally, as expected, ΔT_S obtained from the TR images in Fig. 4(a) is directly proportional to the power dissipation, P [see Fig. 4(b)].

F. Limitation of TR Measurement

Before we close this section, we wish to emphasize some of the fundamental limitations of the TR approach [15]–[17], because metal contacts would one assess the accuracy of the temperature plotted in Section III. Since the positions of the NWs are known *a priori* (because they are defined lithographically), the diffraction limit ($\sim\lambda/2$) manifests alternatively as an uncertainty of the temperature from individual NWs. In particular, the reflectivity of two neighboring NWs (separated by a distance smaller than the diffraction limit) overlaps in such a way that it becomes difficult to assign unique temperature to individual NWs. Moreover, the technique requires careful calibration of $\boldsymbol{k}$, because the reflectivity of the metal surfaces depend on local surface roughness, dictated in turn by the deposition conditions. Finally, it is important to remember that the TR approach measures surface temperature; internal temperature must be measured by complementary techniques [3] or inferred through a theoretical model.

III. Characterization of Self-Heating Transient

In this section, we will use the TR technique to explore several aspects of self-heating in GAA transistors, namely, the spatiotemporal distribution of surface temperature during heating and cooling, the dependence of heating and cooling time constants as a function of the number of NW in a transistor, and so on.

A. Spatiotemporal Temperature Distribution

The high spatiotemporal resolution of TR imaging provides new insights into the transient heating/cooling of a GAA-FET as a function of the number of NW. Fig. 5 shows that during the ON (OFF) state of the V_{DS} pulse, the channel region heats (cools) at 200∼400-ns timescale. The steady-state temperature (ΔT_{SS}) scales with the number of NWs, indicating a significant thermal cross talk among the NWs. Indeed, $\Delta T_{SS} \sim 50K$ at the gate metal surface for a 19-NWs transistor implies even higher self-heating inside the channel, i.e., $\Delta T_C > \Delta T_S$ [3]. Since T_C dictates the performance and the degradation of the transistor, an accurate estimate is desired. Although the methodology to estimate T_C is beyond the scope of this paper, we note in passing that T_C can either be directly measured by complementary experiments, or correlated to the T_S through theoretical modeling as follows.

For the experimental approach, we have previously used the ac output conductance method to independently measure T_C [3]. The results correlate very well with T_S measured by the TR method (see Fig. 6), reflected in the linear increase in T with the number of NW and the fact that T_C is considerably hotter than T_S, as expected. From the theoretical perspective, given the material constants, the thermal compact model

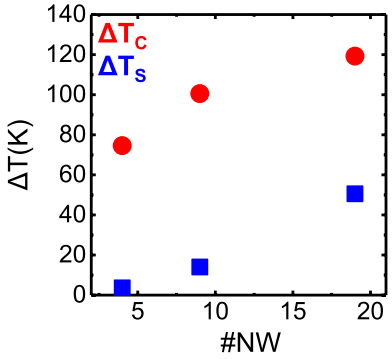

Fig. 6. Increase in the surface temperature and the internal NW channel temperature both scale with the number of NWs. Here, ΔT_S is the peak temperature of the 2-D spatial surface temperature by the TR method described in this paper, and ΔT_C is the average channel temperature measured by the ac output conductance method.

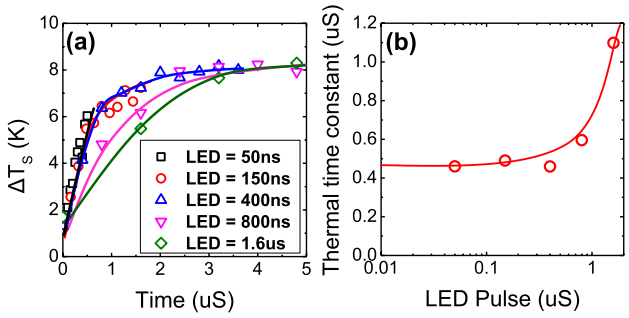

Fig. 7. (a) Transient ΔT_S at the channel surface (Sample B) depending on the LED pulsewidth $\tau_{\rm LED}$ = 50 ns∼1.6 μs. For $\tau_{\rm LED}$ ≤400 ns, the transient profiles overlap, indicating adequate resolution. (b) Saturation of thermal time constants for $\tau_{\rm LED}$ ≤400 ns reflects the overlap of $\Delta T_S(t)$ in (a).

described in Section IV can be used to infer T_C from T_S. A detailed model combining theory and experiments will be discussed in a future paper.

B. Thermal Time Constants

To understand the dynamics of heating/cooling at the operating frequency, it is also important to characterize the time constants for heating and cooling carefully and precisely. To determine the time resolution needed to capture the transient temperature rise, we reduce LED pulse width ($\tau_{\rm LED}$) from 1.6 μs to 50 ns, and check if the heating transients are fully resolved and independent of $\tau_{\rm LED}$. Fig. 7(a) shows the transient heating of the channel surface after $V_{\rm DS}$ pulse is turned ON and characterized with different values of $\tau_{\rm LED}$. The heating transients overlap for $\tau_{\rm LED}$ ≤400 ns, suggesting that $\tau_{\rm LED}$ ∼400 ns provides sufficient temporal resolution. A plot of the effective thermal time-constants, obtained by fitting the heating transients in Fig. 7(a) and summarized in Fig. 7(b), confirms the assertion.

Once the required $\tau_{\rm LED}$ is determined, the transient heating and cooling for GAA transistors with different numbers of NWs are measured [see Fig. 8(a)]. Fig. 8(b) plots the saturated temperature rise ($\Delta T_{\rm SS}$ at t = 0.64 μs) and the thermal time constants as a function of the number of NWs, indicating from the data in Fig. 8(a). We find that the increase in thermal cross talk as a function of the number of NWs increases $\Delta T_{\rm SS}$. The

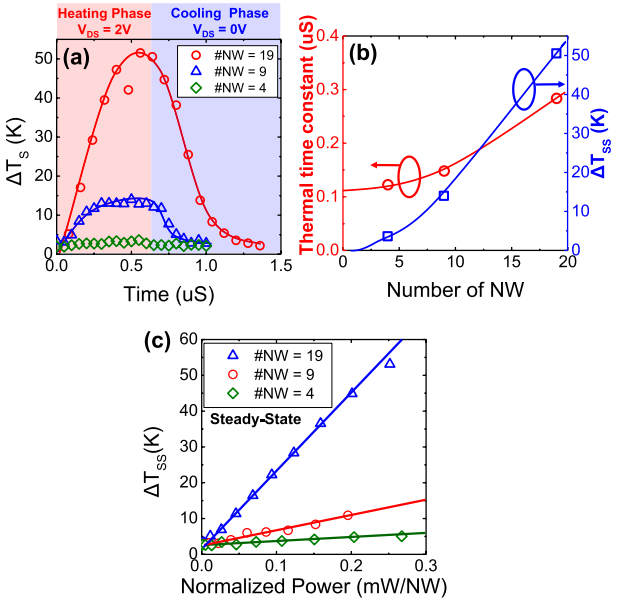

Fig. 8. (a) Transient ΔT_S at the channel surface (Sample A) as a function of the number of NWs as the voltage pulse is applied and then removed. (b) Both ΔT_S (blue square) and the thermal time constants (at 63% of max ΔT_S, red circle) increase with the number of NWs. (c) Both measured ΔT and power ($= V_{DS} \times I_D$) follow similar dependence with drain bias, indicating $\Delta T_S \sim$ power, as expected.

time constants also increase with the number of NWs indicating, that the devices with larger geometry need more time to reach the maximum temperature. Physically, the increase in the time constants reflects the increase effective thermal resistances, confirmed by the increasing slope of the power dissipation versus ΔT_S curves for transistors with different numbers of NWs, as shown in Fig. 8(c). In Section IV, we will use a compact model to explain these empirical observations; the surprising complexity of heat diffusion in these structures will be discussed in a future publication [23].

IV. Interpretation of Self-Heating Time Constant

To understand the experimentally observed features in Section III (e.g., the origin of ∼100-ns time constant), it is important to realize that the GAA transistors involves complex multilayer fabrication. The complexity of the channel geometry and material constants leads to nontraditional heat transport (and multiple time constants) in these structures as follows.

First, note that the aligned array of GAA NWs is covered by Au gate-pad with high thermal conductivity (k_G = 300 W/m/K) [see Fig. 1(a)]. However, since the thermal conductivity of the Atomic Layer Deposition (ALD)-deposited tungsten nitride (WN) gate ($k_{\rm WN}$ = 4 W/m/K) is much smaller than that of the gate oxide ($k_{\rm OX}$ = 50 W/m/K), heat cannot escape directly to the Au-gate; instead, heat must first diffuse through the gate oxide and then exit to the gate contact pad. Unfortunately, heat dissipation through the Au-pad by air convection is not sufficiently fast. Instead, the temperature build-up forces heat to diffuse toward and dissipate through the substrate, Eventually, the substrate acts as the dominant heat sink of the GAA transistor. Despite the complexity of the heat

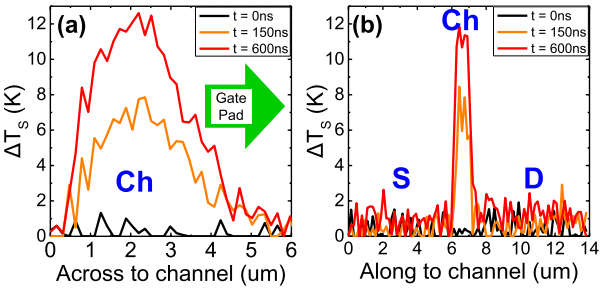

Fig. 9. Temperature plots (a) across and (b) along the channel from Fig. 5 for different times t in heating phase from the nine NWs device.

diffusion, self-heating in these transistors can be explained by three different time constants, associated with three distinct phases of heat diffusion as follows.

For the first phase of self-heating, we may view the NWs to be encapsulated within a thermally insulating surrounding defined by the gate oxide. The relevant time constant for self-heating of the NWs is $\tau_{\rm NW} \sim L_{\rm ox}^{*2}/(k_{\rm OX}/\rho_{\rm OX}C_{V,{\rm ox}}) \sim$1 ns, based on the parameters in Table I. Note that $L_{\rm ox}^{*} = 35$ nm is the effective oxide length to heat flow out to the gate-pad through the oxide. We emphasize that $L_{\rm ox}^{*}$ ($\sim L_{\rm ch}/2$) is not the oxide thickness, because WN gate does not allow direct heat dissipation to the gate. Note that this 1-ns time constant of NW-self heating will not reflected in the TR-based measurement of the surface temperature.

In the second phase, $\Delta T_{\rm NW}^{\rm max}$ is defined by the time needed for the heat to spread over the contact pad. Assuming an effective contact pad width of $W_G^{*} \sim 3$ μm, the effective time constant is given by $\tau_G \sim W_G^{*2}/(k_G/\rho_G C_{V,G}) \sim$100 ns, based on the typical material parameters. This is indeed the time constant observed in Fig. 5, replotted in Fig. 9(a) and (b) for clarity.

In the third phase, the final time constant corresponds to that of the heat dissipation through the substrate, $\tau_{\rm sub} \sim H_{\rm sub}^2/(k_{\rm sub}/\rho_{\rm sub}C_{V-{\rm sub}}) = 60$ ms. Here, $H_{\rm sub} = 350$ μm, $k_{\rm sub} = 3$ W/m/K, $\rho_{\rm sub} = 4810$ Kg/m^3, and $C_{V-{\rm sub}} = 310$ J/Kg/K are the physical thickness, thermal conductivity, mass density, and specific heat, respectively, of the substrate. This slower time constant will be reflected in the surface temperature only for kilohertz heating/cooling transients.

The identification of the three time constants suggests an opportunity to create a simple thermal equivalent circuit Fig. 10(a), which can be used to predict the IC response under a variety of pulse trains as follows. As shown in Fig. 10(b), the input power (P) first heats the GAA-NW, with $\tau_{\rm NW} = 1$ ns. Subsequently, heat spreads all over the highly thermal conducting gate-pad ($\tau_G = 100$ ns). The gate-pad can be viewed as a second reservoir. From the gate-pad, heat flows out through the substrate ($\tau_{\rm sub} = 60$ ms).

The compact model explains the NW-dependence of the steady-state temperature. First, note that if the NWs are thermally isolated (and the substrate and the gate-pad have high thermal conductivities), the temperature rise per NW should be independent of N. In practice, the maximum temperature rise in the NWs, $\Delta T_{\rm NW}^{\rm max} \sim P \times R_{\text{th-NW}} + N \times P \times R_{\text{th-}G}$, where $R_{\text{th-NW}}$ and $R_{\text{th-}G}$ are the thermal resistances

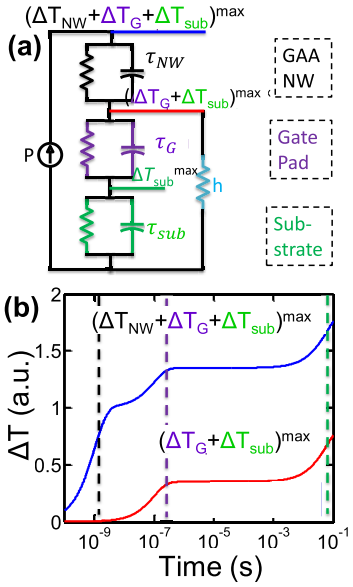

Fig. 10. (a) Thermal circuit model is developed using Foster RC ladder model (Fig. 2a, [24]). (b) Analytical model with realistic thermal time constants ($\tau_{\rm NW} = 1$ ns, $\tau_G = 100$ ns, and $\tau_{sub} = 160$ ms). Maximum NW temperature and maximum surface temperature versus time.

of the each GAA NW and the gate-pad due to lateral heat spreading, respectively. The first component represents the heating in the isolated GAA NW, and the second the heating of the gate-pad by collective power of all NWs. Maximum surface temperature rise follows the gate-pad resistance drop $\Delta T_S^{\rm max} \sim N \times P \times R_{\text{th-}G} \sim N$. Therefore, the surface heats up more with the increase of N. Both τ_G and $\Delta T_S^{\rm max}$ increase almost linearly with N.

V. Thermoreflectance Image of TDDB

Unlike the indirect methods used to date, the high spatiotemporal resolution of TR images may be used to detect the variability and the degradation of individual NW (e.g., $V_{\rm th}$ shift, BD, which impacts the local ON current and consequently local temperature). As an illustrative example, following a gate stress for a certain time (Fig. 11), the channel abruptly becomes very hot, reflecting dielectric BD. Soon thereafter, a few of the NWs are destroyed. With the broken NWs excluded and the temperature of the NW is reduced, $\Delta T_S(x, y)$ of the remaining NWs is restored to pre-BD levels [Fig. 11(d)].

There are several important observations in these images. First, the asymmetric bell-shaped temperature profile in Fig. 11(a) reflects the fact that NWs in the middle have higher self-heating compared with those at the edge—and the heat diffusion toward the right through the gate pad makes the profile asymmetric. Second, the significant increase in temperature between Fig. 11(a) and (b) reflects the difference in the temperature rise early in the self-heating process versus self-heating due to excess gate leakage when the gate dielectric of one of the NWs is broken. The broken NW would also heat the neighboring NWs through thermal cross talk. In time, the significant self-heating is likely to destroy (open) the channel.

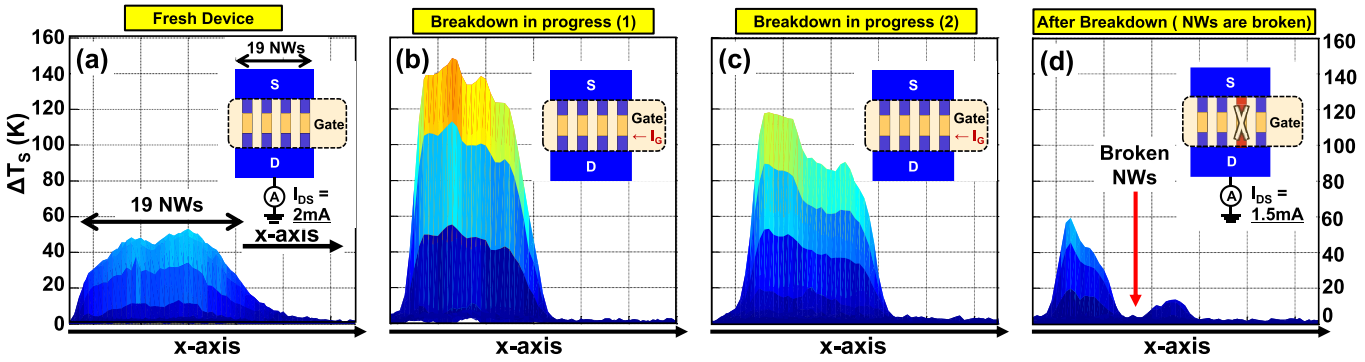

Fig. 11. TR images (side view in Fig. 1, the x-axis is along the width of the channel) at different time instants. After stressing ($V_{DS} = 2$ V, $V_{GS} = 1$ V) Sample A (with 19 NWs) for a certain time, the channel region is suddenly heated due to the increased gate leakage (second image). Eventually, a fraction of the NWs is broken, and the remaining NWs settle the pre-BD temperature (right image). Correspondingly, I_{ON} is decreased from 2 mA at the beginning to 1.5 mA at the end. This clearly indicates that about one-fourth of the NWs are no longer functioning. Inset: schematic of BD of the NWs.

Now, the drain currents in these broken NWs are suppressed; and the channel has been decomposed into two thermally decoupled transistors—on the left with 8–10 NWs, and on the right with 3–5 NWs. The self-heating is reduced significantly, as expected from Fig. 8(c), because some of NWs in the middle are not operational [25]–[27].

This paper demonstrated the detection of BD in NWs level by the TR method. However, the sensitivity of the technique for very low degradation levels and low V_{DS} remain to be established. Note that the temperature change (ΔT_S), as observed in the TR method, is directly proportional to the change of power ($\sim \Delta I$). The TR method can resolve temperature down to 50 mK; therefore, the method should be able to detect a few percentage change in current. For example, if $\Delta T_S \sim 3$ K for nominal I_{ON} at $V_{DS} = 1$ V [Fig. 4(b)], a 2% of change in I_{ON} is expected to produce 60mK change for ΔT_S. One must confirm this assertion through controlled experiments (noise is always a challenge) and is left as a topic of future research.

VI. Conclusion

The high spatiotemporal resolution of the TR imaging offers unprecedented and fundamentally new insights into the mechanics and kinetics of self-heating (e.g., degree of self-heating, dominant heat conduction channel, and dynamics of channel BD) of the emerging multigate technology. Simple thermal compact modeling relates surface temperature with internal temperature, and provides the thermal equivalent circuit for the novel structure. A nuanced use of this versatile technique will help calibrate quasi-ballistic electrothermal modeling tools, assess the relative merits of different multigate topologies, and can eventually improve the cell/circuit layout to suppress SHE as a source of variability and reliability in the modern hyperscaled IC technology.

In particular, regarding the relevance of the technique to other transistor geometries, note that the temperature rise in FinFETs, for example, is much smaller compared with that of the GAA transistors, because the bottom of the fins is directly connected to the substrate, and the gate-stack in FinFET is more thermally conductive than the GAA gate-stack surrounded by poorly conducting ALD-grown WN gate metal. The placement of metal contacts in an industrial FinFET may also not be optimum for TR imaging. These considerations make TR imaging of the standard FinFET more difficult than GAA transistor. Regardless, TR imaging is likely to offer higher spatial and temporal resolution than IR or other imaging techniques. Complemented by other experimental techniques (e.g., ac conductance and IR imaging), TR images can be an important characterization tool for these transistor structures.

Acknowledgment

The authors would like to thank the Birck Nanotechnology Center for the fabrication and characterization facilities. P. D. Ye would like to thank X. Wang and Prof. R. G. Gordon from Harvard University for the technical support in device fabrication.

References

[1] C. Prasad *et al.*, "Self-heat reliability considerations on Intel's 22 nm tri-gate technology," in *Proc. IEEE Int. Rel. Phys. Symp.*, Apr. 2014, pp. 5D.1.1–5D.1.5.

[2] R. Wang *et al.*, "Experimental study on quasi-ballistic transport in silicon nanowire transistors and the impact of self-heating effects," in *IEDM Tech. Dig.*, Dec. 2008, pp. 1–4.

[3] S. H. Shin *et al.*, "Impact of nanowire variability on performance and reliability of gate-all-around III–V MOSFETs," in *IEDM Tech. Dig.*, Dec. 2013, pp. 7.5.1–7.5.4.

[4] L. T. Su, J. E. Chung, D. A. Antoniadis, K. E. Goodson, and M. I. Flik, "Measurement and modeling of self-heating in SOI nMOSFET's," *IEEE Trans. Electron Devices*, vol. 41, no. 1, pp. 69–75, Jan. 1994.

[5] B. M. Tenbroek, M. S. L. Lee, W. Redman-White, R. J. T. Bunyan, and M. J. Uren, "Self-heating effects in SOI MOSFETs and their measurement by small signal conductance techniques," *IEEE Trans. Electron Devices*, vol. 43, no. 12, pp. 2240–2248, Dec. 1996.

[6] R. H. Tu, C. Wann, J. C. King, P. K. Ko, and C. Hu, "An AC conductance technique for measuring self-heating in SOI MOSFET's," *IEEE Electron Device Lett.*, vol. 16, no. 2, pp. 67–69, Feb. 1995.

[7] N. Beppu, S. Oda, and K. Uchida, "Experimental study of self-heating effect (SHE) in SOI MOSFETs: Accurate understanding of temperatures during AC conductance measurement, proposals of 2ω method and modified pulsed IV," in *IEDM Tech. Dig.*, Dec. 2012, pp. 28.2.1–28.2.4.

[8] K. A. Jenkins, J. Y.-C. Sun, and J. Gautier, "Characteristics of SOI FET's under pulsed conditions," *IEEE Trans. Electron Devices*, vol. 44, no. 11, pp. 1923–1930, Nov. 1997.

[9] X. Wang, S. Farsiu, P. Milanfar, and A. Shakouri, "Power trace: An efficient method for extracting the power dissipation profile in an IC chip from its temperature map," *IEEE Trans. Compon. Packag. Technol.*, vol. 32, no. 2, pp. 309–316, Jun. 2009.

[10] J. McDonald and G. Albright, "Microthermal imaging in the infrared," *Electron. Cooling*, vol. 3, no. l, pp. 26–29, Jan. 1997.

[11] D. D. Griffin, "Infrared techniques for measuring temperature and related phenomena of microcircuits," *Appl. Opt.*, vol. 7, no. 9, pp. 1749–1756, Sep. 1968.

[12] A. Sarua *et al.*, "Integrated micro-Raman/infrared thermography probe for monitoring of self-heating in AlGaN/GaN transistor structures," *IEEE Trans. Electron Devices*, vol. 53, no. 10, pp. 2438–2447, Oct. 2006.

[13] D. L. Blackburn, "Temperature measurements of semiconductor devices—A review," in *Proc. 20th Annu. IEEE Semiconductor Thermal Meas. Manage. Symp.*, Mar. 2004, pp. 70–80.

[14] A. Shakouri, K. Maize, P. Jackson, X. Wang, B. Vermeersch, and K. Yazawa, "Ultrafast submicron thermal characterization of integrated circuits," in *Proc. 19th IEEE Int. Symp. Phys. Failure Anal. Integr. Circuits (IPFA)*, Jul. 2012, pp. 1–2.

[15] J. Christofferson, K. Maize, Y. Ezzahri, J. Shabani, X. Wang, and A. Shakouri, "Microscale and nanoscale thermal characterization techniques," *J. Electron. Packag.*, vol. 130, no. 4, pp. 041101-1–041101-6, 2008.

[16] M. Farzaneh *et al.*, "CCD-based thermoreflectance microscopy: Principles and applications," *J. Phys. D, Appl. Phys.*, vol. 42, no. 14, pp. 143001-1–143001-20, 2009.

[17] K. Maize, E. Heller, D. Dorsey, and A. Shakouri, "Fast transient thermoreflectance CCD imaging of pulsed self heating in AlGaN/GaN power transistors," in *Proc. IEEE Int. Rel. Phys. Symp.*, Apr. 2013, pp. CD.2.1–CD.2.3.

[18] J. J. Gu, Y. Q. Liu, Y. Q. Wu, R. Colby, R. G. Gordon, and P. D. Ye, "First experimental demonstration of gate-all-around III–V MOSFETs by top-down approach," in *IEDM Tech. Dig.*, Dec. 2011, pp. 33.2.1–33.2.4.

[19] J. J. Gu *et al.*, "20–80 nm channel length InGaAs gate-all-around nanowire MOSFETs with EOT = 1.2 nm and lowest SS = 63 mV/dec," in *IEDM Tech. Dig.*, Dec. 2012, pp. 27.6.1–27.6.4.

[20] S. Shin *et al.*, "Origin and implications of hot carrier degradation of gate-all-around nanowire III–V MOSFETs," in *Proc. IEEE Int. Rel. Phys. Symp.*, Jun. 2014, pp. 4A.3.1–4A.3.6.

[21] Z. M. Zhang, B. K. Tsai, and G. Machin, Eds., *Radiometric Temperature Measurements: II. Applications* (Experimental Methods in the Physical Sciences), vol. 43. New York, NY, USA: Academic, 2009.

[22] S. Dilhaire, S. Grauby, and W. Claeys, "Calibration procedure for temperature measurements by thermoreflectance under high magnification conditions," *Appl. Phys. Lett.*, vol. 84, no. 5, pp. 822–824, 2004.

[23] M. A. Wahab, S. H. Shin, and M. A. Alam, "Analysis of self-heating in 3D transistor," *IEEE Trans. Electron Devices*, to be published.

[24] M. N. Touzelbaev, J. Miler, Y. Yang, G. Refai-Ahmed, and K. E. Goodson, "High-Efficiency Transient Temperature Calculations for Applications in Dynamic Thermal Management of Electronic Devices," *J. Electron. Packag.*, vol. 135, no. 3, p. 031001, Jul. 2013.

[25] M. A. Alam and R. K. Smith, "A phenomenological theory of correlated multiple soft-breakdown events in ultra-thin gate dielectrics," in *Proc. IEEE Int. Rel. Phys. Symp.*, Mar./Apr. 2003, pp. 406–411.

[26] M. A. Alam, B. E. Weir, and P. J. Silverman, "A study of soft and hard breakdown—Part I: Analysis of statistical percolation conductance," *IEEE Trans. Electron Devices*, vol. 49, no. 2, pp. 232–238, Feb. 2002.

[27] M. A. Alam, B. E. Weir, and P. J. Silverman, "A study of soft and hard breakdown—Part II: Principles of area, thickness, and voltage scaling," *IEEE Trans. Electron Devices*, vol. 49, no. 2, pp. 239–246, Feb. 2002.

Authors' photographs and biographies not available at the time of publication.