Direct Observation of Self-heating in III-V Gate-all-around Nanowire MOSFETs

S. H. Shin*, M. Masuduzzaman, M. A. Wahab, K. Maize, J. J. Gu, M. Si,
A. Shakouri, P. D. Ye, and M. A. Alam*

*E-mail: {shin136, alam}@purdue.edu, Phone: (765) 494-5988, Fax: (765)-494-2706
Department of ECE, Purdue University, West Lafayette, IN 47907, USA

Abstract

Gate-all-around MOSFETs use multiple nanowires to achieve target I_{ON}, along with excellent 3D electrostatic control of the channel. Although self-heating effect (SHE) has been a persistent concern, the existing characterization methods, based on indirect measure of mobility and specialized test structures, do not offer adequate spatio-temporal resolution. In this paper, we develop an *ultra-fast, high resolution* thermo-reflectance (TR) imaging technique to (*i*) directly observe the increase in local surface temperature of the GAA-FET with different number of nanowires (NWs), (*ii*) characterize/interpret the time constants of heating and cooling through high resolution transient measurements, (*iii*) identify critical paths for heat dissipation, and (*iv*) detect in-situ time-dependent breakdown of individual NW. Our approach also allows indirect imaging of quasi-ballistic transport and corresponding drain/source asymmetry of self-heating. Combined with the complementary approaches that probe the internal temperature of the NW, the TR-images offer a high resolution map of self-heating in the surround-gate devices with unprecedented precision, necessary for validation of electro-thermal models and optimization of devices and circuits.

Introduction

Multi-gate devices, such as, FinFET, Gate-all-around transistors (GAA-FET) improve 3D electrostatic control of the channel, but the corresponding increase in self-heating may compromise both performance and reliability. Although the self-heating effect of FinFET appears significant, but tolerable [1], the same may not be true for GAA geometry [2, 3], especially in quasi-ballistic regime where hot spots and non-classical heat-dissipation pathways may lead to localized heating and damage. The existing reports of the SHE on the SOI, FinFET or GAA-FET have so far relied either on indirect electrical measurements such as AC output conductance and gate resistance [4, 5] with inherent temporal delays, or on optical infra-red (λ >1.5um) imaging that cannot resolve deep sub-micron features so that it requires customized large structure. As a result, although the SHE has become a critical issue in GAA-FET, it has so far been impossible to fully resolve the spatio-temporal features of the SHE.

In this paper, we first develop an *ultra-fast, high resolution* thermo-reflectance (TR) imaging technique to directly observe the local time-dependent rise in the surface temperature, $\Delta T(x, y, t)$. A variety of devices with different number of nanowires (NW) and oxide thicknesses are explored, and the effect of these parameters on the self-heating are interpreted and analyzed. For example, high resolution transient measurements allow us to characterize the time constants of heating and cooling of the channel, define surround-gate oxides as the primary heat conduction pathways for thermal dissipation, and interpret NW-specific self-heating and degradation of the transistor.

Experimental Setup

The devices used in this study are InGaAs GAA n-MOSFET (Fig. 1), with different oxides thicknesses (T_{ox}), channel lengths (L_{ch}), and # of NWs. The fabrication process is described in [6, 7] and the device dimensions are listed in Table 1. The steep subthreshold slope, reported experimentally in [6, 7], is reproduced by the 3D Sentaurus simulation (Fig. 2), confirming excellent electrostatic control on GAA-FET [8].

During the TR imaging [9, 10], a high-speed LED ($\lambda =$ 530nm) pulse illuminates the device, and a synchronized CCD camera captures the reflected image with ~ 250nm spatial resolution (Fig. 3). Theoretically, this technique relies on the change of the complex refractive index of a material with differential increase in temperature(ΔT), so that the change in local reflectance of the device surface relates to ΔT as,

$$\frac{\Delta R}{R_0} = \frac{1}{R_0} \cdot \left.\frac{dR}{dT}\right|_{T=T_0} \Delta T \equiv \boldsymbol{k} \cdot \Delta T$$

where $\boldsymbol{k}$ (K^{-1}) is the thermoreflectance coefficient [10]. The calibration of $\boldsymbol{k}$ allows a CCD image to be interpreted as a map of $\Delta T(x, y)$, with 50mK resolution. For the transient measurement of $\Delta T(x, y, t)$, the device is periodically turned on and off by V_{DS} pulse (Fig. 3b) 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 ~ 50ns resolution. As a basic validation, Fig. 4 shows that $\Delta T \propto$ Power, as expected.

Characterization of Self-Heating

The high spatio-temporal resolution of TR imaging provides new insights into the transient heating/cooling of a GAA-FET as a function of #NW. Fig. 5 shows that during the ON (OFF) state of the V_{DS} pulse, the channel region heats (cools) at ~ 200ns timescale. The steady state temperature (ΔT_{SS}) scales with the #NW, indicating significant thermal cross-talk among the NWs. Indeed, ΔT_{SS} ~ 50K at the gate metal surface for a 19-NWs transistor implies even higher self-heating inside the channel [2].

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 (τ_{LED}) from 1.6us to 50ns, and check if the heating transients are fully resolved and independent of τ_{LED}. Fig. 6a shows the transient heating of the channel surface after V_{DS} pulse is turned ON and characterized with different τ_{LED}. The heating transients overlap for $\tau_{LED} \leq$ 400ns, suggesting that $\tau_{LED}\sim$400ns provides sufficient temporal resolution. A plot of the effective thermal time-constants, obtained by fitting the heating transients in Fig. 6a and summarized in Fig. 6b, confirms the assertion. Once the required $\tau_{LED}$is determined, the transient heating and cooling for transistors with different number of NWs are measured (Fig. 7a). Fig. 7b shows that the increase in thermal cross-talk with the # of NWs increases saturated ΔT_{SS}. The time constants also increase with #NW 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 resistance, confirmed by the increasing slope of the power dissipation vs. ΔT curves for transistors with different #NW (Fig. 8).

Optimization of Self-Heating and Reliability Measurements

Identification of Heat Conduction Channel: The ON current can be improved by reducing SHE; this requires identification and subsequent optimization of the heat conducting channels. To find the primary heat conduction channel among the substrate, source-drain contacts, or the gate contacts (see Fig. 1), we measured self-heating for transistors with different T_{ox} (Fig. 9a). We find that the transistor wrapped with thicker oxide (Al_2O_3) shows reduced SHE, indicating dominant heat flow along S/D: the thicker oxide offers *lower* thermal resistance to S/D (Fig. 9b); the opposite would have been true if substrate or gate channels were dominant. A solution of the heat equation in the relevant geometry confirms this assertion that the heat flux escapes laterally to the S/D contacts along the oxide itself (Fig. 9c) (note that Al_2O_3 has higher thermal conductivity ($= 35W/(m \cdot K)$) than the immediate gate metal-WN). The premise is also supported by the observation of increased temperature on S/D contact metal in Fig. 10. In order to see clear ΔT_{SS} change depending on power dissipation, V_{DS} is varied from 0 to 4V under same V_{GS} = 1V. As expected, the asymmetry of heating near drain side (vs. source) reflects asymmetry in heat generation in quasi-ballistic sub-100nm GAA transistor. The spatial extent of heating in the drain pad (~1um), marked should not be misinterpreted as being due to direct heating by the ballistic electrons. Instead, the ballistic electrons are likely to act as a heat source at the drain-edge (~100nm) [8]; the subsequent diffusion of heat in the metal is observed by the TR-approach.

Reliability Measurements: Unlike the indirect methods used to date, the high spatio-temporal resolution of TR images can be used to detect the variability and degradation of individual NW (e.g. V_{th} shift, breakdown, which impacts the local ON current and consequently local temperature). As an illustrative example, following a gate stress for certain time (Fig. 11), the channel abruptly becomes very hot, reflecting dielectric BD and thermal cross-talk among the NWs. Soon thereafter, a few of the NWs are destroyed. With the broken NWs excluded and the cross-talk suppressed, $\Delta T(x, y)$ of the remaining NWs is restored to pre-BD levels (Fig. 11 (right)).

Conclusions

The high spatio-temporal 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, dynamics of channel breakdown) of the emerging multi-gate technology. A nuanced use of this versatile technique will help calibrate quasi-ballistic electro-thermal modeling tools, assess the relative merits of different multi-gate topologies, and can eventually improve the cell/circuit layout to suppress SHE as a source of variability and reliability in modern hyper-scaled IC technology.

Acknowledgement

We acknowledge Birck Nanotechnology Center for the fabrication and characterization facilities. Prof. Ye thanks Xinwei Wang and Prof. Roy G. Gordon from Harvard University for the technical support in device fabrication.

References

[1] S. Ramey, A. Ashutosh, C. Auth, J. Clifford, M. Hattendorf et al., in IEEE International Reliability Physics Symposium (IRPS), 2013 pp. 4C.5.1-4C.5.5.

[2] S. H. Shin, M. Masuduzzaman, J. J. Gu, M. A. Wahab, N. Conrad, M. Si, P. D. Ye, M. A. Alam, in IEEE International Electron Device Meeting (IEDM), 2013 pp. 7.5.1-7.5.4.

[3] R. Wang, J. Zhuge, C. Liu, R. Huang, D. W. Kim, D. Park, Y. Wang, in IEEE International Electron Device Meeting (IEDM), 2008 pp. 753-756.

[4] W. Jin, W. Liu, Samuel K. H. Fung, Philip. C. H. Chan, Chenming Hu, IEEE Transactions on Electron Devices, vol.48, no.4, pp.730-736, April 2001.

[5] K. Jenkins, J. Sun, J. Gautier, IEEE Transactions on Electron Devices, vol.44, no.11, pp.1923-1930, Nov 1997.

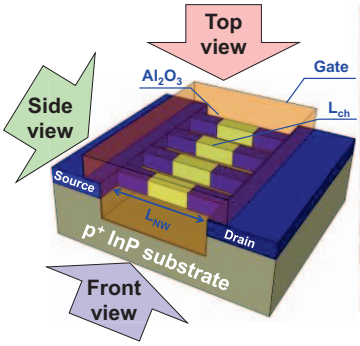
[6] J. J. Gu, Y. Q. Liu, Y. Q. Wu, R. Colby, R. G. Gordon, and P. D. Ye, in IEEE International Electron Device Meeting (IEDM), 2011 pp. 769-772.

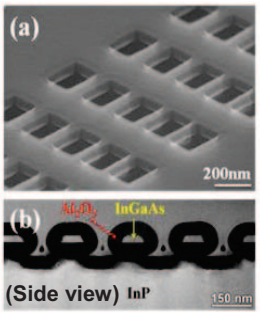
[7] J. J. Gu, X. W. Wang, H. Wu, J. Shao, A. T. Neal, M. J. Manfra, R. G. Gordon, P. D. Ye, in IEEE International Electron Device Meeting (IEDM), 2012 pp. 633-636.

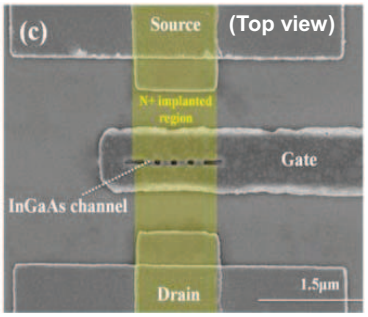
[8] S. H. Shin, M. A. Wahab, M. Masuduzzaman, M. Si, J. J. Gu, P. D. Ye, M. A. Alam, in IEEE International Reliability Physics Symposium (IRPS), 2014 pp. 4A.3.1-4A.3.6.

[9] K. Maize, E. Heller, D. Dorsey, A. Shakouri, in IEEE International Reliability Physics Symposium (IRPS), 2013 pp. CD.2.1-CD.2.3.

[10] Zhang, Radiometric Temperature Measurements, Volume 43: II. Applications (Experimental Methods in the Physical Sciences), Academic Press, 2009.







	Sample A IEDM 2011 [6]	Sample B IEDM 2012 [7]
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	10nm Al_2O_3	3.5nm Al_2O_3
EOT (nm)	4.5	1.7
# of NW	1, 4, 9, 19	4

Fig. 1: (left) Schematic image of an InGaAs GAA NW n-channel MOSFET. (a) SEM image of parallel NWs. (b) STEM image of the cross section of the InGaAs NWs (Side view). (c) SEM image of the parallel InGaAs NWs (Top view). Images taken from Ref. [6].

Table. 1: The description of the two types of samples (different T_{ox} and different number of the NWs, L_{ch} = 70~80nm, and W_{NW} = 30nm) used in this study.

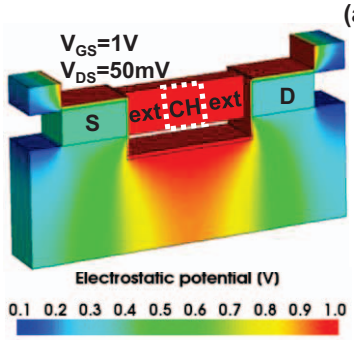


Fig. 2: Simulated potential profile of the GAA MOSFET (Sample A) for V_{GS} = 1V and V_{DS} = 50mV. Strong gate controllability over the nanowires is confirmed [8].

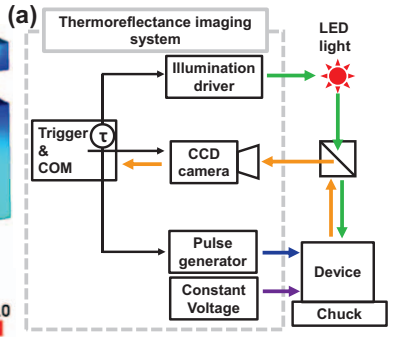


Fig. 3: (a) Schematics of thermoreflectance (TR) imaging system. A pulse generator (V_{DS}) and a constant voltage source (V_{DS}) drive the transistor. A control computer triggers the illumination driver and the CCD camera for a given delay time with respect to V_{DS}. (b) The timing diagram for transient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

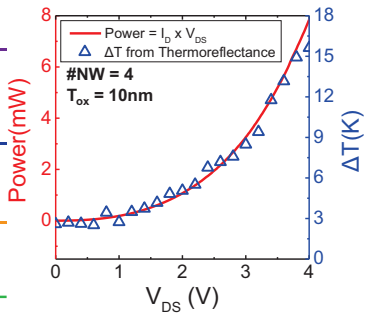


Fig. 4: Both measured ΔT and power ($= V_{DS} \times I_D$) follow similar dependence with drain bias, indicating $\Delta T \propto Power$, as expected.

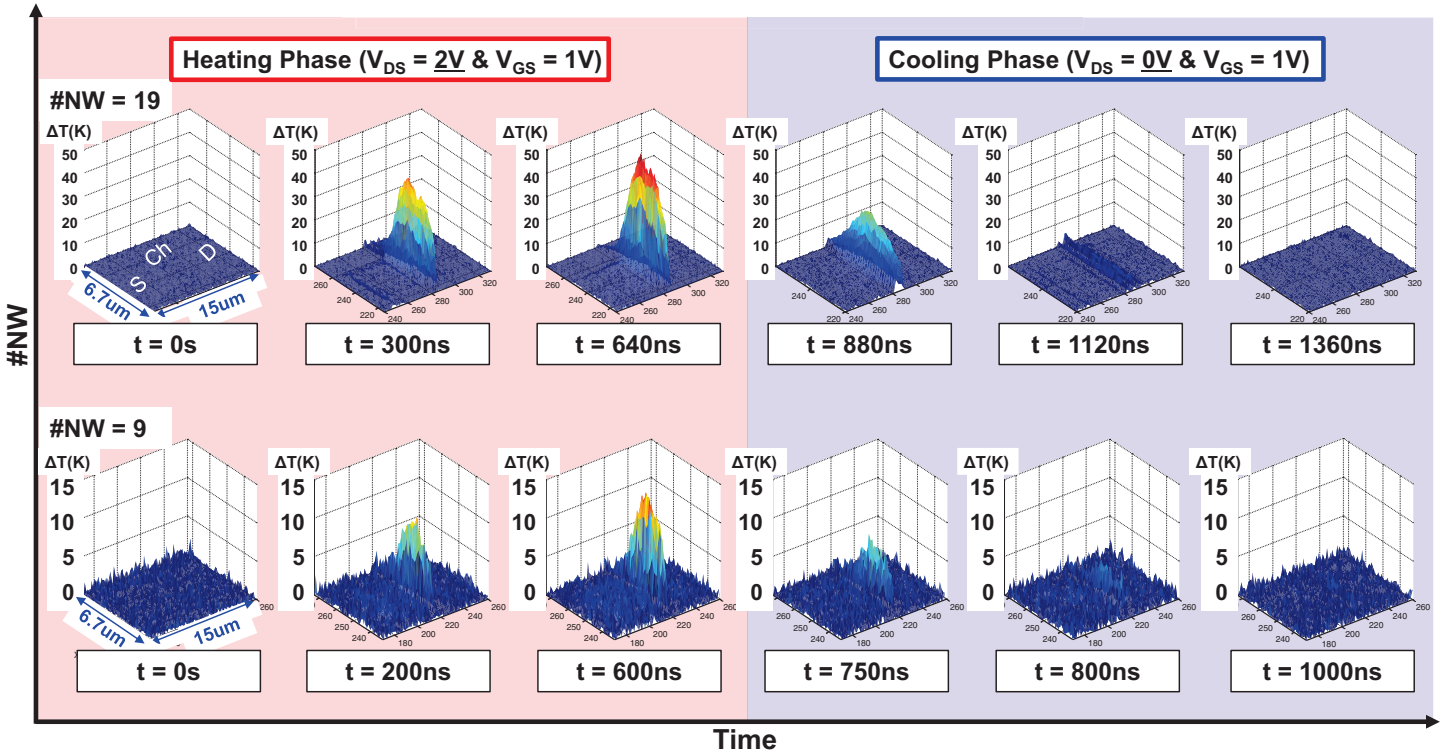


Fig. 5: Three dimensional TR images for heating (V_{DS} = 2V & V_{GS} = 1V) and cooling (V_{DS} = 0V & V_{GS} = 1V) phases. For clarity, only 3 images (out of more than 15) per cycle per device are shown. Also, images of 4 NWs are available, but not shown. The device with 19 NWs (top) shows higher saturation temperature compared to the device with 9 NWs (bottom). The heating and cooling time constants lie on the order of 100-500ns, depending on #NW, oxide thickness, etc.

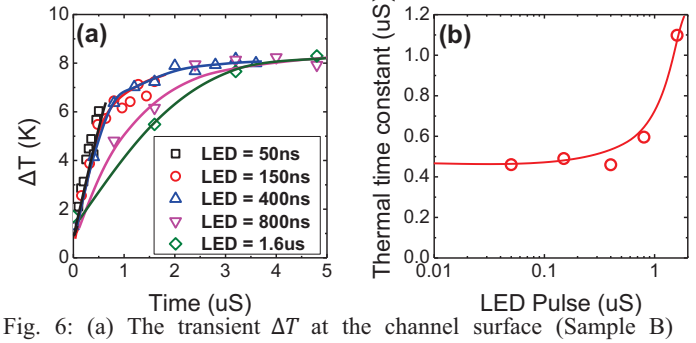


Fig. 6: (a) The transient ΔT at the channel surface (Sample B) depending on the LED pulse width (τ_{LED} = 50ns ~ 1.6us). For τ_{LED} ≤ 400ns, the transient profiles overlap, indicating adequate resolution. (b) The saturation of thermal time constants for τ_{LED} ≤ 400ns reflects the overlap of $\Delta T(t)$ in part (a).

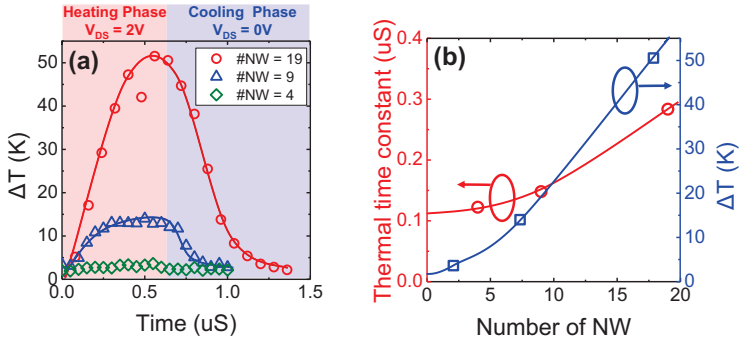


Fig. 7: (a) The transient ΔT at the channel surface (Sample A) as a function of the #NWs as the voltage pulse is applied and then removed. (b) Both ΔT (blue square) and the thermal time constants (at 63% of max ΔT, red circle) increase with the number of NWs.

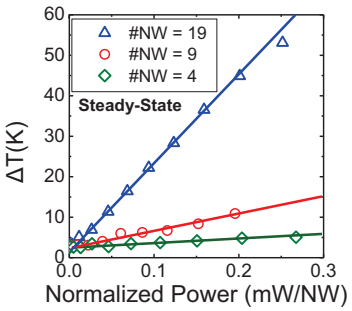


Fig. 8: ΔT increases linearly with the normalized power dissipation per NW. However, devices with higher number of NWs show higher ΔT for a given normalized power, indicating higher effective thermal resistance.

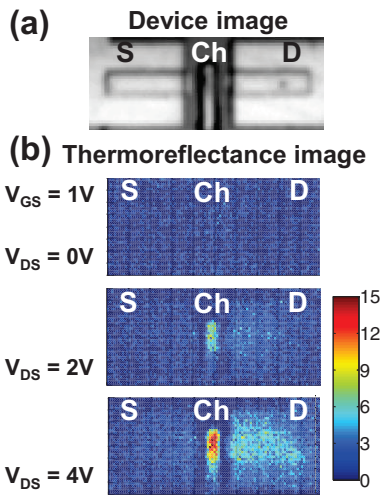


Fig. 10: (a) Device image (4 NWs for Sample A) from top view. (b) TR images from top view by varying V_{DS} = 0 ~ 4V and fixed V_{GS} = 1V. We observe that not only the gate is heated, but also, source and drain pads are heated due to heat flows through the oxide layer from heat source at drain-edge.

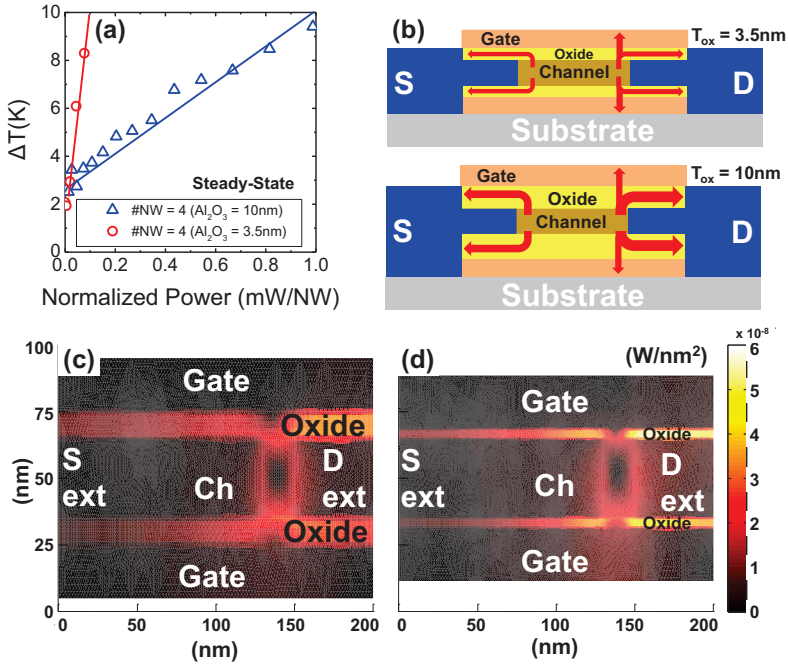


Fig. 9 (Simulation): (a) For given dissipation power per NW, the temperature rise is much higher for a thinner oxide (T_{ox} = 3.5nm) as compared to a thicker oxide (T_{ox} = 10nm). (b) For thicker oxide heat can flow more easily to the source/drain contacts, as shown schematically. (c, d) Simulation of heat flux (W/nm²) for T_{ox}=10nm (c) and 3.5nm (d) indicates higher heat flow through the oxide, due to the relatively higher thermal conductivity of Al_2O_3 compared to the first layer of gate metal (WN).

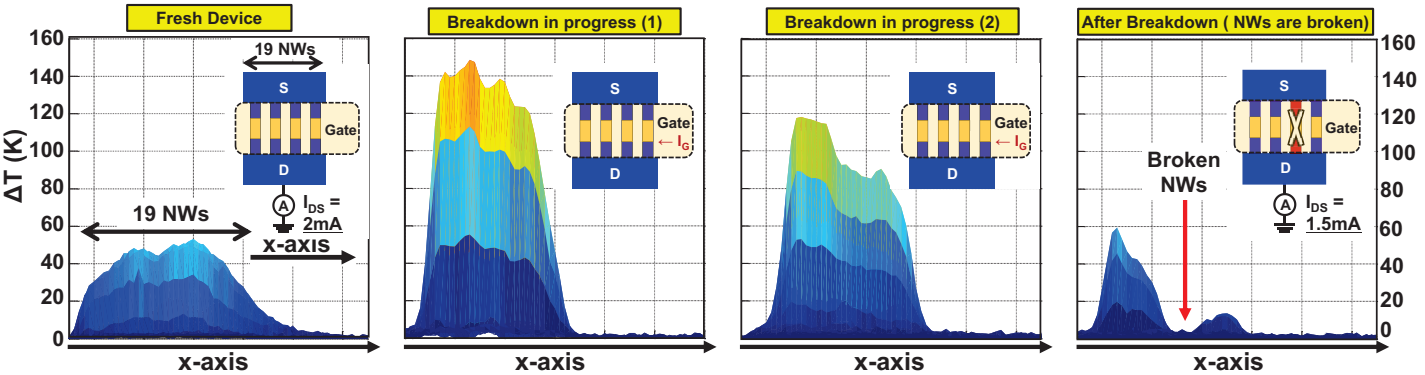


Fig. 11: TR images (Side view in Fig. 1, x-axis is along the width of the channel) at different time instants. After stressing (V_{DS} = 2V, V_{GS} = 1V) Sample A (with 19 NWs) for certain time, the channel region is suddenly heated due to the increased gate leakage (2nd image). Eventually, a fraction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