Impact of Self-Heating on Performance, Power and Reliability in FinFET Technology

Victor M. van Santen, Paul R. Genssler, Om Prakash, Simon Thomann, Jörg Henkel and Hussam Amrouch
Department of Computer Science, Karlsruhe Institute of Technology, Karlsruhe, Germany
{victor.santen, genssler, om.prakash, simon.thomann, henkel, amrouch}@kit.edu

(Invited Paper)

***Abstract*—Self-heating is one of the biggest threats to reliability in current and advanced CMOS technologies like FinFET and Nanowire, respectively. Encapsulating the channel with the gate dielectric improved electrostatics, but also thermally insulates the channel resulting in elevated channel temperatures as the generated heat is trapped within the channel. Elevated channel temperatures lowers the performance, increases leakage power and degrades the reliability of circuits. Self-heating becomes worse in each new transistor structure (from planar transistor to FinFET to Nanowire) due to the ever-increasing thermal resistance of the transistor. This leads to elevated temperatures, which must be carefully considered while designing circuits. Otherwise, reliability cannot be ensured. This work presents a self-heating study to illustrate how self-heating matters in digital circuits. It also explores the impact of running workloads in SRAM arrays, such as register files in CPUs, and how self-heating effects in SRAM cells can be mitigated.**

***Index Terms*—Self-Heating, Temperature, Reliability, FinFET, Nanowire, SRAM**

I. Introduction

Temperature is a well-known reliability concern for semiconductor devices. In the following, we first present an overview of the general impact of temperature on the performance, power and reliability of circuits. Then, we illustrate the impact of self-heating, that results in elevated temperatures inside transistors. To explain Self-Heating Effects (SHE), we cover the technology trend from planar MOSFET to FinFET and finally Nanowire transistors, discussing where the heat comes from and what it ultimately leads to. This links SHE to temperature effects. To explain SHE, we start first with a high-level overview of the impact of temperature in general.

A. Impact of Temperature on Transistors

When temperature rises, the operation of a transistor changes in two key parameters [1]: 1) The carrier mobility μ of the transistor decreases. 2) The threshold voltage V_{th} decreases. From the perspective of circuit designers, this alters the drain current I_D of a transistor. The drain current I_D in the ON-state of the transistor is called I_{ON} and in the OFF-state I_{OFF}. When the transistor heats up, the decrease in V_{th} increases I_{ON}, but the decrease in μ decreases it. For regular supply voltages V_{DD}, the decrease in mobility has a stronger impact on I_{ON}, than impact of the V_{th} decrease and hence I_{ON} reduces as a result. For I_{OFF}, the lower V_{th} is the dominant factor and thus I_{OFF} exponentially increases at higher temperatures. In short, the hotter a transistor, the higher I_{OFF} (i.e., higher leakage power) and the lower I_{ON} (i.e., larger propagation delays and thus lower performance).

B. Impact of Temperature on Circuit Performance

As a circuit designer, the circuit metrics must be derived from these changes in the transistor properties while operating at elevated temperatures. Performance in a circuit is governed by the time a transistor can charge (discharge) its load capacitance C_{load} (e.g., transistor gates of the following circuitry). When I_{ON} drops due to elevated temperatures, the time to charge the capacitance increases and thus the propagation delay t_{delay} of a circuit increases.

The temperature-induced performance loss can be modeled in circuit simulations with the industry-standard transistor compact model (i.e., BSIM [2]). The transistor model describes the transistor properties based on the temperature. Then, circuit SPICE simulations estimate the impact of the lower I_{ON} and higher I_{OFF} on the propagation delay of the circuit as well as on its static power.

The impact of elevated temperatures on circuit's performance can be reduced mitigated using temperature-aware logic synthesis. In [3], we showed how to design a circuit with smaller performance loss at elevated temperatures in a fully automated manner by employing temperature-aware cell libraries. In [4], we demonstrated how to employ approximate computing concepts to reduce narrow timing guardbands required to protect against temperature effects.

C. Impact of Temperature on Circuit Power Consumption

The power consumption of a CMOS circuit can be divided in the static power consumption P_{static} and the dynamic power consumption P_{dyn}. The static power consumption is the sum of all the leakage currents (i.e., I_{OFF}) in the circuit. Since, I_{OFF} increases, the power consumption of the circuit increases. This leads to the famous leakage/static power feedback loop, in which an increase in temperature leads to an increase in static power, which leads to an increase in temperature and so on. This positive feedback can lead to the thermal runaway problem if not accounted for.

At the same time, since I_{ON} decreases, P_{dyn} decreases. Yet, since the circuit now needs longer (higher t_{delay}) to charge the same C_{load}, the total energy consumption goes up.

978-1-7281-4123-7/20/$31.00

To reduce the impact of temperature on power, we can dynamically scale the voltage down as in [5]. Reducing V_{DD} reduces both P_{dyn} and P_{static} and thus reduces the power overhead introduced by the temperature.

D. Impact of Temperature on Circuit Reliability

Temperature can affect the reliability of circuits in two key ways. First, it directly reduces I_{ON} and thus shrinks timing guardbands due to longer t_{delay}. This is directly reversible by lowering the temperature and contrary to aging, has no history or accumulations over time. As such, it can be solved by closely monitoring, at runtime, the available guardband or with including, at design-time, accurate timing guardbands.

Secondly and more importantly, temperature is a strong stimulus for transistor aging effects. Bias Temperature Instability (BTI) and Hot-Carrier Injection (HCI) are currently the dominant aging effects and both are accelerated by temperature increase, as shown in [6] [7]. Accelerated aging increases V_{th} further, leading to more I_{ON} reduction and hence more performance losses. Prolonging t_{delay} thus either reduces performance (lowering f_{clk} as a response) or results in timing violations (if f_{clk} is not dynamically lowered) and thus harms the functionality of the circuit. Even worse, since aging occurs non-uniformly in a circuit due to thermal gradients across the chip, it can exhibit the actual worst-case condition. We have shown in [8], how the actual worst-case is worse than uniform peak degradations in each transistor, due to transistor interactions. Some transistors provide negative feedback to circuit operation (e.g., leakage in pull-up PMOS transistors counteracts the NMOS transistor in an inverter, when the inverter tries to pull down the voltage at its output) and thus reach the worst-case t_{delay} when the counteracting transistors are fresh ($\Delta V_{th} = 0$) instead of at peak degradation.

The impact of aging effects on the circuit is not solely deterministically. In fact, since aging is stochastic in nature, the variability of the circuit increases, as shown in [9]. This is critical since aging-induced variability already dominates overall variability [10] and therefore any further stimulation of aging must be limited as much as possible to contain the reliability problems.

II. Self-Heating exacerbates Temperature Impact

So far, we have discussed temperature, in general, which meant the temperature of the chip T_{chip}. To explain SHE, we additionally need to introduce T_C, which is the temperature of the transistors channel. In this section, we show how with each new generation of semiconductor technology, SHE becomes worse. SHE is not a new phenomenon, it started as a negligible phenomenon in planar transistors and then, over time, turned into one of the most critical threats to CMOS reliability in current FinFET transistor structures and future Nanowire transistor structures.

A. Introduction to Self-Heating in Transistors

The channel of a transistor heats up due to a high current flowing through a resistive channel. This creates heat due to Joule heating [11]. This heat flux raises the channel temperature T_C and worsens this particular transistor as discussed in Section I-A.

1) Self-Heating in MOSFET: In planar MOSFET, SHE is already present but not an issue, since the channel can conduct its heat to the substrate below the channel (low thermal resistance R_{th} between channel and substrate), which is cooled via the cooling of the chip. Therefore, despite the existence of SHE in planar MOSFET, it is negligible, as $\Delta T_C \approx 0$. However, in high-power MOSFET, the Joule heating is strong enough to elevate T_C due to very high currents and voltages, despite decent thermal conductance (low R_{th}) to the substrate. Consequently, SHE can appear $\Delta T_C > 0$ in planar technologies if the joule heating is sufficiently large, like in high-power MOSFET [12].

2) Self-Heating in SOI-MOSFET: Unfortunately, planar MOSFETs could not be scaled indefinitely as I_{OFF} became too high. In order to improve I_{OFF}, transistor designers started to electrically insulate the transistors channel. The first step was Silicon-On-Insulator (SOI), which insulated the channel from the circuit with a buried oxide layer. Now, parasitic diodes in the substrate were removed, lowering I_{OFF}. However, this also results in limited thermal conduction (high R_{th}) from the channel to the substrate and SHE became much higher (elevated T_C) [11].

3) Self-Heating in FinFET: To further continue scaling, the channel control had to be improved. For this purpose, instead of applying the electric field solely vertically, the 2D planar MOSFET was changed into a 3D FinFET structure. In FinFETs the channel is surrounded on three sides by the gate, which provides stronger electric field and thus more control over the channel. However, now the gate encapsulates the channel in both an electrical and thermal insulator (the gate dielectric). This further increases R_{th} and thus T_C in FinFET compared to planar MOSFET [13].

To illustrate the magnitude of SHE in FinFET, we employed a TCAD simulation of a 14 nm FinFET transistor calibrated with data from Intel [14]. This simulation allows us to study the resulting high T_C across the entire 3D volume of the transistor, which is shown in Fig. 1. SHE can reach $T_C = 100°C$ near the drain of the transistor and at the top of the fin, because from there R_{th} is the highest towards the substrate. The skew towards the drain is due to heat transport via the electrons flowing through the channel, which results in a heat flux towards the drain of the transistor.

Importantly, SOI-FinFET are required for 7nm transistors and beyond, which increases R_{th} and thus T_C even further [13]. Now, heat can only be conducted through the metal source and drain contacts as all other 4 sides (top, bottom, left, right) are now thermal insulators (gate dielectric and buried oxide in substrate).

FinFETs grow in discrete steps, when more current is required than a single fin can provide. While planar MOSFET just increase their width to provide higher I_D, FinFET use 1, 2, 3, etc. fins. When these multi-fin FinFET are used, SHE becomes even worse [13] due to the larger current densities

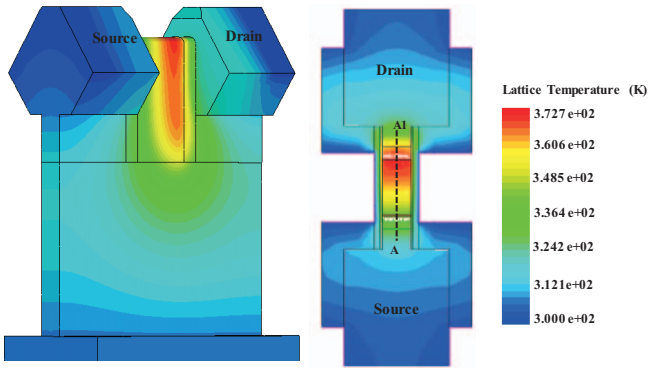


Fig. 1. 14nm FinFET transistor exhibiting high channel temperature T_C.

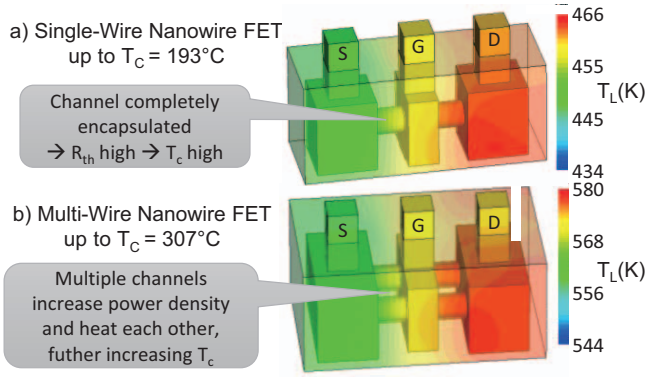


Fig. 2. a) Nanowire with single wire (channel) exhibiting very high T_C b) Nanowire with multiple wires (channels) exhibiting extraordinary T_C [17].

and confined structure. For example, in a 3-fin transistor, the fin itself is heated via the source and drain metal contacts and the heat sink of the drain (the substrate) is also heated by its neighboring fins. Therefore, it experiences heating and a higher R_{th}. In our work in [15], we have modeled the fin-dependence of SHE in FinFET transistors in the standard BSIM-CMG transistor model [16].

4) Self-Heating in Nanowires: The next step in semiconductor technology is to fully encapsulate the channel, i.e. Gate-All-Around (GAA) or Nanowire transistors. In these Nanowire transistors, the channel is completely surrounded with electrically and thermally insulating gate dielectric. This provides improved channel control compared to FinFET. However, R_{th} is incredibly high and the channel reaches extraordinarily high T_C due to SHE. Even worse, when multiple wires are used (similar to multi-fin FinFET), the individual wires heat each other (either via the gate dielectric or drain/source metal contacts) and increase the power density in this tiny volume. Therefore, multi-wire Nanowire transistors feature the highest SHE-induced T_C, which can reach $> 200°C$ in our detailed TCAD simulations shown in Fig. 2 taken from [17].

B. Self-Heating and Temperature Effects

As we have seen in Section I, elevated temperature degrades transistors and thus has a harmful impact on circuit performance, power and reliability. For the operation of a circuit, T_{chip} or $T_{ambient}$ are not actually the temperatures, that really matter. In fact, the flow of carriers is controlled in the channel of the transistor and thus T_C is the temperature, which ultimately governs the operation (i.e., the electrical properties) of a transistor. As presented in Fig. 1, T_C for FinFET can reach $100°C$, while for Nanowires in Fig. 2 up to $> 200°C$ can be observed in TCAD simulations [17]. Any ΔT_C due to SHE is on top of the baseline $T_{ambient}$ range from $25°C$ to $125°C$ resulting in the final T_C. For instance, the $T_C = 300°C$ Nanowires shown in Fig. 2 consist of $125°C$ $T_{ambient}$ and SHE-induced $\Delta T_C \approx 180°C$.

C. Traditional Thermal Management cannot mitigate Self-Heating Effects in Transistors

Since SHE is a thermal problem, intuitively, the circuit designer might resort to thermal management techniques. However, traditional thermal management techniques are not effective for SHE. For example, increasing the cooling lowers the temperature of T_{chip}, but this barely has an impact on T_C. The source of SHE is the thermal insulation (high R_{th}) of the channel from its surroundings, thus lowering T_{chip} to lower T_C cannot work.

Also other techniques like Dynamic Voltage and Frequency Scaling (DVFS) cannot mitigate SHE. While SHE itself depends on Joule heating and thus I_D, which is lowered by lowering V_{DD}, this only lowers SHE-induced ΔT_C if we *permanently* lower V_{DD}. However, DVFS tries to exploit the thermal capacitance on an chip or fluctuations in activity to lower heating (by lowering V_{DD}) if T_{chip} reaches or exceeds $T_{critical}$. Traditionally, DVFS works great, as the chip needs a couple of seconds to heat up or cool down (due to the thermal capacitance of heat spreader, heat sink, etc.). Therefore, lowering V_{DD} only when necessary provides an increase in performance. Yet, the thermal capacitance in SHE are orders of magnitude smaller, since now only the thermal capacitance of the channel matters, not the entire chip and its cooling system (heat spreader, heat sink, etc.). SHE heats up and cools in a couple of nanoseconds [18] [11], compared to seconds for the chip. Since this is in the order of single clock cycles, DVFS cannot react swiftly enough and cannot reduce SHE-induced T_C.

III. Design for Self-Heating

Since Self-Heating cannot be mitigated with traditional techniques, two options emerge. Either, the structure of the transistor is changed in our favor, reducing R_{th}, or the circuits have to be hardened to cope with high T_C. The first option is not feasible, since technology trends are clearly heading in the opposite direction. Electrical control over the channel became paramount in smaller geometries, pushing us to SOI, FinFET and finally Nanowires, increasing R_{th} in each step. Therefore, our only option is to harden the circuits to sustain operation even under ever-fluctuating and high T_C. To harden the transistors against high T_C, we need to accurately estimate SHE-induced ΔT_C.

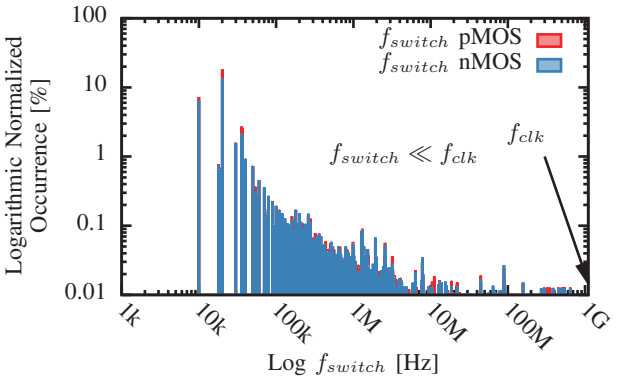


Fig. 3. Distribution of f_{switch} across transistors in a Processor [15]. This highlights that $f_{switch} \ll f_{clk}$ and thus SHE in digital circuits matters.

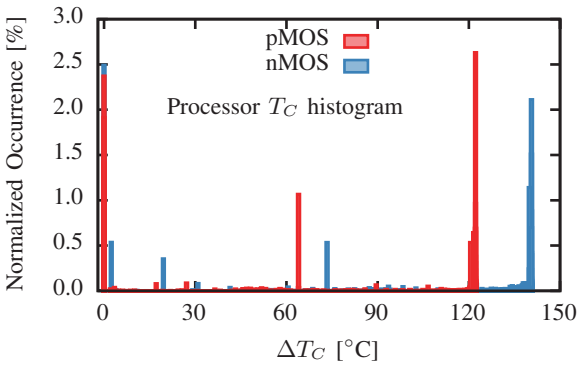


Fig. 4. Distribution of T_C across transistors in a processor [15]

A. Estimate Self-Heating

To estimate SHE, we must estimate T_C in each transistor or provide an upper bound for T_C for all transistors. The latter would provide then an upper bound for timing, which can be used to create cell libraries (compare [8]). Constant heating due to a constant current flowing through a transistor is simple upper bound for T_C, but as seen in Fig. 1 and 2, this can reach very high T_C and thus very pessimistic designs.

1) Estimate Transistor Switching: Since constant current is too pessimistic, T_C must be estimated based on the activities of the circuit. Activity in this context means how frequently a transistor switches f_{switch}. This f_{switch} determines how frequently current is flowing through the transistors and thus T_C (SHE is essentially a low pass filter [19] [18]).

Traditionally, SHE was considered negligible in digital circuits, since they operate at $f_{clk} > 1GHz$ and thus have heating and cooling cycles alternating so quickly, that T_C would just saturate at a low steady-state temperature [19]. However, in our recent work [15] we have shown that $f_{switch} \ll f_{clk}$ with some instances of $f_{switch} \approx 10kHz$ despite operating a processor at $f_{clk} = 1.1GHz$ (see Fig. 3). Thus, digital circuit do indeed suffer from SHE, as their f_{switch} is much lower than originally anticipated. In [15], we have presented an approach to efficiently extract the f_{switch} and duty cycle (on-/off-ratio) of a transistor, for every transistor in a processor. Therefore, to consider SHE accurately, one could take the activity of every transistor and then simulate T_C. Since this requires a SPICE simulation of the entire processor, this is not computationally feasible for large circuits. Therefore, we explore upper bounds in Section III-D.

B. Self-Heating Model

BSIM-CMG the FinFET compact transistor model [16], which is widely used in circuit simulations (e.g., in SPICE), supports SHE. It models SHE via three parallel circuit components, exploiting the duality between temperature and electricity. The first is a current source representing heat flux, which has a value equal to the power loss in the transistor $I_{power} \equiv P_{loss} = I_D \cdot V_{DS}$. Next is a capacitor C_{th} equal to the thermal capacitance of the transistor. Finally, a resistor R_{th}, which represents the thermal resistance from the channel to its surroundings (substrate, etc.). These three components form a low-pass filter with the time constant $\tau_{th} = C_{th} \cdot R_{th}$, which is in the order of nanoseconds. Therefore, high frequency activity (high f_{switch}) barely affects ΔT_C, while low frequency activity create high shifts in T_C.

C. Thermal Dependence of R_{th}

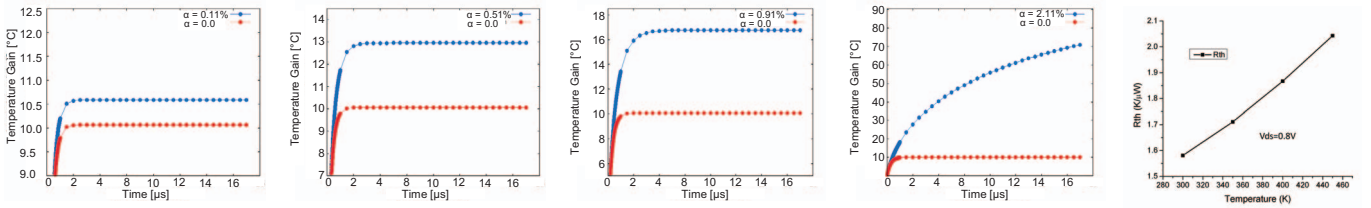
The thermal resistance R_{th} of the gate dielectric is a material property, which itself is temperature dependent. When R_{th} changes, it affects SHE modeling, so we investigated if this temperature dependency should be considered. We have used TCAD simulations of a 14nm FinFET transistor to calculate the change in thermal resistance R_{th} over different temperatures (shown in Fig. 5e). Then, the thermal dependency of the thermal resistance can be expressed with the factor α in the following simplistic equation (purely to explore the dependency, accurate modeling should be more sophisticated):

$$R_{th} = 1 + \alpha \cdot T_C \tag{1}$$

Our results (shown in Fig. 5) indicate that unless $\alpha > 0.5\%$, $\Delta T_C < 3°C$ and thus does not have to be considered. Using our data from Fig. 5e, we estimate $\alpha = 0.18\%$ and thus $\Delta T_C < 1°C$ (compare to Fig. 5a). Therefore, in current technologies it is not necessary to consider the thermal dependency of thermal resistance of the gate dielectric.

D. Worst-Case Self-Heating

In Section III-A1, we have briefly discussed how the activity per transistor can be extracted. The extraction presented in [15] is computationally feasible (couple of hours), but to simulate each transistor under SHE is computationally not feasible in a design process, where the design is optimized iteratively (nevertheless shown in Fig. 4). Thus, an upper bound would be much better to simply simulate the worst-case instead of every transistor within a processor. Instead of constant heating, it is possible to find a less pessimistic worst case with the data from [15]. This worst case estimates SHE for the lowest occurring frequency, which in this case would be $f_{switch} \approx 10kHz$ (see Fig. 3). Additionally, a duty cycle is necessary to know how much of the $0.1ms$ period ($10kHz \rightarrow t_{period} = 0.1ms$) the transistor is on (heating) and off (cooling). Then, with lowest



(a) $\alpha = 0.11\%$ $\Delta T_C < 0.5°C$ (b) $\alpha = 0.51\%$ $\Delta T_C > 3°C$ (c) $\alpha = 0.91\%$ $\Delta T_C > 6°C$ (d) $\alpha = 2.11\%$ $\Delta T_C > 60°C$ (e) 14nm FinFET TCAD simulation [14] exhibits slope $\alpha = 0.18\%$

Fig. 5. Different α in Eq. 1 show an impact on T_C if $\alpha > 2\%$ in a SPICE simulation of a 14 nm transistor at $V_{DD} = 0.7V$ with $f_{switch} = 1GHz$. Since α in current technologies is about 0.18% (see Fig. 5e), it is unnecessary to consider the thermal dependence of R_{th}. Note scale in Y-axis.

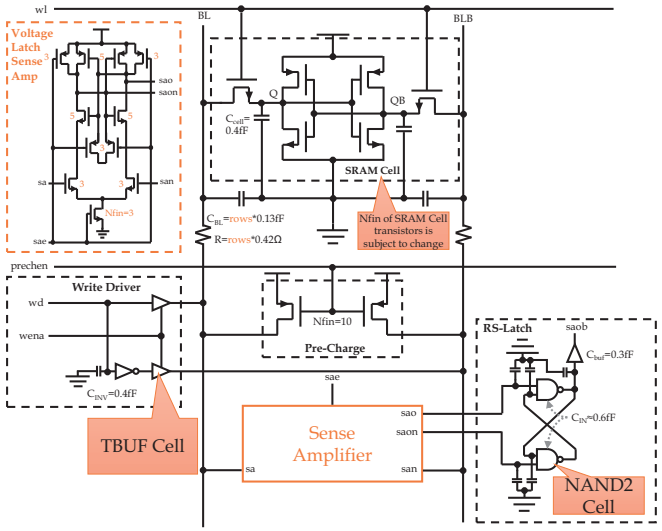


Fig. 6. Schematic of the simulated SRAM Array

f_{switch} and highest duty cycle, worst SHE-induced ΔT_C is estimated as an upper bound for that workload (application running on the processor) on that circuit. Since, this is the lowest actually occurring frequency with highest occurring duty cycle it must have the highest ΔT_C peaks (assuming the highest voltage is identical with V_{DD} for all transistors) and thus provides a safe upper bound. The bound still entails some pessimism, but is far off from constant current simulations with constant heating.

IV. Self-Heating in Circuits

To study the impact of SHE in circuit, various circuits are used. Most studies use ring oscillators [11] [13] or other simple circuits to illustrate the impact of SHE. However, with their regular patterns and simple activity (ring oscillators have one f_{switch} and uniform duty cycle of 0.5), they cannot capture the impact SHE has on larger circuits. Some studies for larger circuits exist, like our own previous work in [7] which studied an entire SRAM array (SRAM cells with latched Sense Amplifier (SA), Write Driver (WD) and pre-charging circuit) under SHE. The schematic of the circuit is shown in Fig. 6. In this work, we want to build on top of our work in [7] and introduce activity from real processors. Previously, we studied solely abstract activity for the cell and focused on the addition of periphery (SA, WD, etc.), while in this work, we can explore the actual activity and different design options for SRAM arrays with respect to SHE.

To obtain the transistor activity, circuit activity needs to be extract. Thus read, write and hold operations are extracted from a gem5 processor simulation [20] to simulate ARMv8 CPU cores and monitor their register file represented with a 32 $\times$ 64-bit SRAM array. Different applications result in different activities for the cells, SA and WD. These different activities result in different temperatures in different transistors as visible in Fig. 7a. A more than $60°C$ difference in temperature can be observed in the difference between these two curves as shown in Fig. 7b. This clearly indicates, that while estimating SHE in circuit, their workload (circuit activity such as applications for processors, images for image processing, etc.) must be taken into account.

The origin of this workload dependency is given by the f_{switch} and duty cycle of the transistors. In a SRAM array, the access transistors are conducting, when the word line WL is high, i.e. the cell is written or read. Similarly, the SA is only active during a read in its column or the WD is only active during a write. These metrics are entirely driven by the application, which might read all the time, but seldom write or vice versa. The values being written govern which side of the SRAM cell, SA and WD are charging (discharging) the bit lines. With respect to SHE multiple read and write operations are solely important if these read (writes) are consecutive. If the operations are not consecutive, then the transistor may have already cooled down (due to the low τ_{th}). The values read or written matter as well. If the operations feature opposite values (read 0 $\rightarrow$ read 1), the heated transistor performing the operation have sufficient time to cool down. If the operations read the same value (read 0 $\rightarrow$ read 0, etc.) then the transistor remain hot and additional heat is added, further increasing SHE-induced ΔT_C.

Besides the workload, the design of the SRAM array itself is also important. Our register file has 32 rows, i.e. the bit lines are short (low bit line capacitance C_{BL}) and low number of cells per column. However, in caches the number of rows is higher due to the less stringent performance constraints and much higher memory capacity. This has a large impact on the observed ΔT_C, since now the WD and cell have to work against much larger C_{BL} to charge (discharge). Fig. 8 shows peak ΔT_C of the access transistors (pass-gate PG) and NMOS transistors (pull-down PD) of the SRAM cell in the array

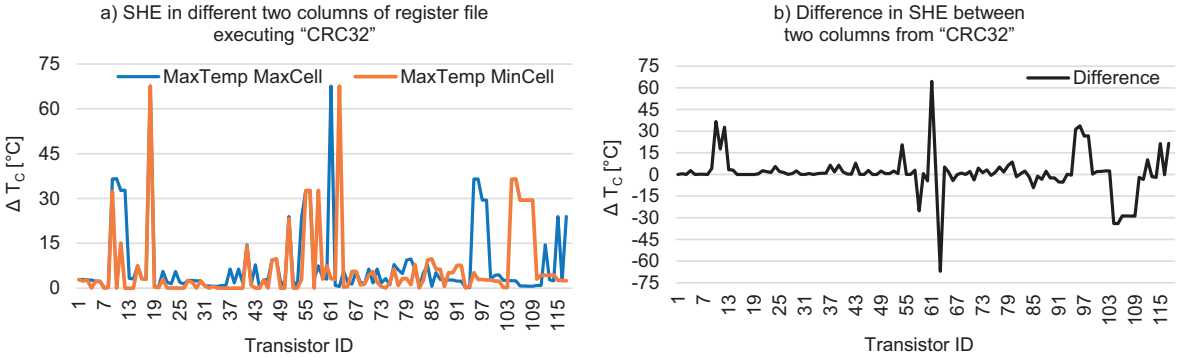


Fig. 7. Maximum SHE-induced ΔT_C according to a SPICE simulation of all 119 transistors in schematic (Fig. 6) featuring cells from 2 different columns with different read, write and hold patterns (not many reads in “MinCell” and lots of reads in “MaxCell”). SPICE settings are $V_{DD} = 0.8V$, $f_{clk} = 3GHz$.

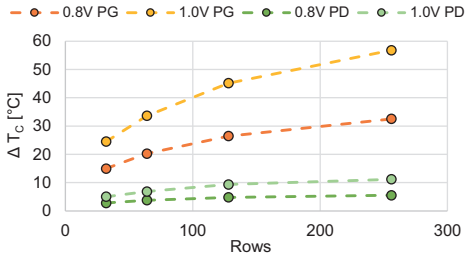


Fig. 8. Impact of number of rows (bit line capacitance) per periphery (SA and WD) on peak SHE-induced ΔT_C in SRAM arrays. For 1.0V PG the difference between 32 and 256 rows (8x higher C_{BL}) is more than 2x higher ΔT_C. PG: Pass-Gate transistor and PD: Pull-Down transistor in SRAM cell.

during the read and write operations. A large difference in ΔT_C can be observed in Fig. 8 if the number of rows in an array grows. More rows increases C_{BL}, which increases the duration a cell is written or read, thus increasing ΔT_C of the access transistors. Similarly, the PD transistors have to resist a write longer if the write itself takes longer due to higher C_{BL}, increasing their ΔT_C. In summary, the design of the SRAM array, especially the number of rows per periphery (sometimes called memory bank) matters with respect to SHE.

V. Conclusion

This work provides an overview over self-heating in circuits. We discussed how self-heating got worse in each new transistor technology and presented multiple techniques on how to design for self-heating based on the activities within the circuit. Our results illustrate that during the design process self-heating-induced temperatures must be estimated based on the workload of the circuit.

Acknowledgement

We thank Yogesh S. Chauhan, Chetan K. Dabhi and Souvik Mahapatra for their support in our TCAD simulation. Additionally, we thank Michael Meinschäfer for his support in the exploration of the R_{th} temperature dependence.

References

[1] D. Wolpert and P. Ampadu, “Temperature effects in semiconductors,” in *Managing temperature effects in nanoscale adaptive systems*. Springer, 2012, pp. 15–33.

[2] B. Sheu, D. Scharfetter, and P.-K. Ko et. al., “BSIM: Berkeley short-channel IGFET model for MOS transistors,” *JSSC*, 1987.

[3] H. Amrouch, B. Khaleghi, and J. Henkel, “Optimizing temperature guardbands,” in *Design, Automation Test in Europe Conference Exhibition (DATE), 2017*, March 2017, pp. 175–180.

[4] B. Boroujerdian, H. Amrouch, J. Henkel, and A. Gerstlauer, “Trading off temperature guardbands via adaptive approximations,” in *2018 IEEE 36th International Conference on Computer Design (ICCD)*, Oct 2018, pp. 202–209.

[5] H. Amrouch, B. Khaleghi, and J. Henkel, “Voltage adaptation under temperature variation,” in *2018 15th International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD)*. IEEE, 2018, pp. 57–60.

[6] V. M. van Santen, J. Martin-Martinez, H. Amrouch, M. M. Nafria *et al.*, “Reliability in super- and near-threshold computing: A unified model of rtn, bti, and pv,” *TCAS-I*, 2018.

[7] H. Amrouch, V. M. v. Santen, O. Prakash, H. Kattan *et al.*, “Reliability challenges with self-heating and aging in finfet technology,” in *2019 IEEE 25th International Symposium on On-Line Testing and Robust System Design (IOLTS)*, July 2019, pp. 68–71.

[8] V. M. van Santen, H. Amrouch, and J. Henkel, “New worst-case timing for standard cells under aging effects,” *IEEE Transactions on Device and Materials Reliability*, vol. 19, no. 1, pp. 149–158, 2019.

[9] V. M. van Santen, H. Amrouch, and J. Henkel, “Modeling and mitigating time-dependent variability from the physical level to the circuit level,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, pp. 1–14, 2019.

[10] X. Wang, A. R. Brown, B. Cheng, and A. Asenov, “Statistical variability and reliability in nanoscale FinFETs,” in *IEDM*, 2011.

[11] W. Ahn, S. Shin, C. Jiang, H. Jiang *et al.*, “Integrated modeling of self-heating of confined geometry (finfet, nwfet, and nshfet) transistors and its implications for the reliability of sub-20nm modern integrated circuits,” *Micro. Rel.*, 2018.

[12] C. Anghel, A. Ionescu, N. Hefyene, and R. Gillon, “Self-heating characterization and extraction method for thermal resistance and capacitance in high voltage mosfets,” in *ESSDERC’03. 33rd Conference on European Solid-State Device Research, 2003.* IEEE, 2003, pp. 449–452.

[13] D. Jang, E. Bury, R. Ritzenthaler, M. G. Bardon *et al.*, “Self-heating on bulk finfet from 14nm down to 7nm node,” in *IEDM*, 2015.

[14] S. Mishra, H. Amrouch, J. Joe, C. K. Dabhi *et al.*, “A simulation study of nbti impact on 14-nm node finfet technology for logic applications: Device degradation to circuit-level interaction,” *IEEE Transactions on Electron Devices*, vol. 66, no. 1, pp. 271–278, Jan 2019.

[15] V. M. van Santen, H. Amrouch, and J. Henkel, “On the workload dependence of self-heating in finfet circuits,” *TCAS-II*, 2019.

[16] M. V. Dunga, C.-H. Lin, A. M. Niknejad, and C. Hu, “Bsim-cmg: A compact model for multi-gate transistors,” in *FinFETs and Other Multi-Gate Transistors*. Springer, 2008, pp. 113–153.

[17] O. Prakash, S. Manhas, J. Henkel, and H. Amrouch, “Impact of NBTI Aging on Self-Heating in Nanowire FET,” in *DATE*, 2020.

[18] S. Makovejev, S. Olsen, and J. Raskin, “Rf extraction of self-heating effects in finfets,” *TED*, 2011.

[19] H. Jiang, S. Shin, X. Liu, X. Zhang *et al.*, “The impact of self-heating on hci reliability in high-performance digital circuits,” *LED*, 2017.

[20] N. Binkert, B. Beckmann, G. Black, S. K. Reinhardt *et al.*, “The Gem5 Simulator,” *SIGARCH Comput. Archit. News*, 2011.