Impact of Self-heating on Performance and Reliability in FinFET and GAAFET Designs

Vidya A. Chhabria and Sachin S. Sapatnekar
Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455
Email: {chhab011, sachin}@umn.edu

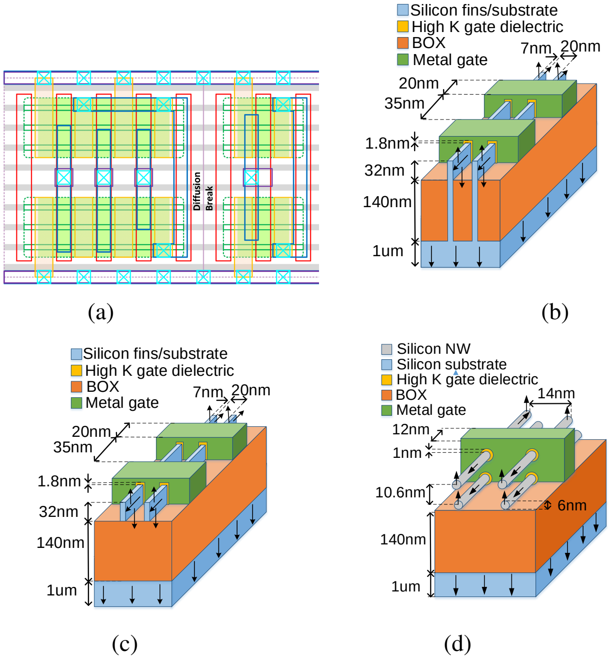
***Abstract*—Modern transistors such as FinFETs and gate-all-around FETs (GAAFETs) suffer from excessive heat confinement due to their small size and three-dimensional geometries, with limited paths to the thermal ambient. This results in device self-heating, which can reduce speed, increase leakage, and accelerate aging. This paper characterizes the temperature for both the 7nm FinFET and 5nm GAAFET sub-structures and analyzes its impact on circuit performance (delay and power) and reliability (bias temperature instability, hot carrier injection, and electromigration). On average, logic gates in a circuit heat up by 12K for 7nm SOI FinFET and by 17K for 5nm GAAFET designs. This rise in temperature accelerates delay degradation due to bias temperature instability and hot carrier injection by up to 25% in FinFET and 39% in GAAFET designs, and also degrades the electromigration-induced time to failure of wires by up to 38% in SOI FinFET and 45% in GAAFET technologies.**

I. Introduction

Traditional planar MOS devices at the 28nm node and higher have been built as (i) bulk MOSFETs on bulk Si wafers, or (ii) silicon-on-insulator (SOI) MOSFETs built above an insulating buried oxide (BOX) layer that improves performance by reducing leakage and parasitics. To enable efficient scaling, designs at the 16/14nm node are based on multigate 3D FinFETs that provide improved electrostatic control over the channel. These device topologies help reduce short channel effects, increase the drive current, enable the use of lower supply voltages, and provide superior scalability. These structures may also be constructed as bulk FinFETs on a bulk substrate, or SOI FinFETs, built above a BOX layer. The SOI FinFET provides similar advantages over the bulk FinFET as the SOI MOSFET over its bulk counterpart. Transistors continue to evolve to further enhance the gate surface area, while shrinking the device footprint. While the FinFET covers three surfaces of the fin, the gate-all-around FET (GAAFET) completely surrounds the channel. Lateral GAAFETs, with vertically stacked silicon nanowires (NWs), could replace FinFETs at the 5nm node. Vertical GAAFETs with vertical NWs can be extremely scalable, and are predicted to be used beyond 5nm.

FinFETs and GAAFETs are susceptible to self-heating (SH). The top-level view of a standard cell in the ASAP7 FinFET technology [1] is shown in Fig. 1(a). The high transistor density implies that a great deal of heat is dissipated per unit footprint. A cross-sectional view shows how the heat generated within the channel is transferred to the ambient through the substrate and the metal interconnects. Thus, the low thermal conductivity of oxide insulators, and low fin/gate pitches, result in high heat flux with restrictive paths to the ambient.

Fig. 1(b) shows the heat transfer paths in a 7nm bulk FinFET structure based on [1]. In the SOI FinFET (Fig. 1(c)), the thick BOX layer with low thermal conductivity degrades the effectiveness of heat conduction through the substrate, and heat flows through the wires to the ambient too. A 5nm lateral GAAFET (Fig. 1(d), based on [2]) has horizontal NWs in pillars, surrounded by oxide, with difficult paths to the ambient through the substrate; a significant amount of heat flows through the metal interconnects. In all these structures SH can hurt circuit performance and accelerate aging degradation [3], which typically worsens exponentially with temperature.



(a) (b)

(c) (d)

Figure 1: (a) Adjacent NAND3 and INV cells in the ASAP7 library [1]. Structure and the paths of heat dissipation in (b) 7nm bulk FinFET, (c) 7nm SOI FinFET, and (d) 5nm lateral GAAFET with arrows that indicate the paths to thermal ground.

Prior works on SH largely focus on a single device or a gate with simplified assumptions on the thermal environment [4], e.g., setting source/drain contacts to the ambient, or neglecting heat flow through the BOX layer (shown not to be true in [3]), or using empirical approaches [5]. Given the exponential thermal dependence of reliability mechanisms, these errors could incorrectly predict the impact of bias temperature instability (BTI), hot carrier injection (HCI), and electromigration (EM).

Our work builds an accurate finite difference based thermal model for circuits built using FinFETs and GAAFETs. The model provides precise estimates of the device temperatures through detailed modeling. We first develop the model at the transistor-level and then use the principle of superposition to estimate the temperature of structures with multiple gates and fins/NWs. The power distributions within the channel of the transistor, between multiple fins, and gates of the cell have been meticulously accounted for, based on their probabilities of switching. We develop a thermal estimation methodology that leverages layout regularity in advanced nodes (e.g., Fig. 1(a) shows that gates are laid out as arrays of fins and gate terminals). Based on this property, we characterize a look-up table (LUT) that enables accurate and rapid thermal analysis of FinFET/GAAFET arrays/substructures. Finally, we embed this gate-level analysis into circuit-level timing analysis for logic blocks, and accurately analyze the impact of temperature on performance, aging, and on EM-constrained circuit lifetime.

II. Self-heating Model

A. Finite Difference Method to Solve 3-D Heat Equation

The second order partial differential equation that governs the conduction of heat in three-dimensional space, (x, y, z), as shown in, is [6]:

$$C\frac{\partial T}{\partial t} = K\left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} + P(x, y, z, t)\right) \qquad (1)$$

where K is the thermal conductivity (W/mK), $T(x, y, z, t)$ is the temperature, $P(x, y, z, t)$ is the rate of heat generation (W/m^3), t is time (s), and C is the volumetric heat capacity (J/(m^{3}K)).

For long-term analysis of the impact of SH on reliability, analyzing thermal behavior in the steady state is appropriate. Under this condition, $\frac{\partial T}{\partial t}$ is zero, and based on the finite differences method (FDM) [6], the electrical-thermal equivalence can be used to build a thermal resistance network. The power dissipation (or heat) in each element is modeled as a current source. If there are n finite regions, then the temperature at each FDM node can be obtained by solving a system of n linear equations:

$$G\mathbf{T} = \mathbf{P} \qquad (2)$$

Here, $G \in \mathbb{R}^{n\times n}$ is the thermal conductance matrix, $\mathbf{T}, \mathbf{P} \in \mathbb{R}^n$ are vectors of the unknown temperatures, and the power dissipations within each element, respectively. The G matrix can be formed in $O(n)$ time using standard nodal analysis, and the values in G depend on thermal conductivities between two neighboring elements and the boundary conditions. A standard linear equation solver can be used to solve this system.

B. Thermal Analysis of Advanced FETs

The values in $\mathbf{P}$ are determined by the distribution of the dissipated power within the device. To a first order, they can be obtained from SPICE-level simulations of unit structures, such as FinFET arrays or logic gates. However, while such circuit simulation models provide the power dissipation per transistor, they do not describe the distribution of power *within* the transistor. Our thermal simulation uses a discretization that places multiple FDM cells within each transistor, and a more fine-grained distribution of power is necessary.

In the confined region within a fin or a Si NW, the high energy electrons that carry current scatter with the lattice vibrations (phonons) which cause the lattice to heat up. This phenomenon reduces the mobility of the carriers and hence reduces the average phonon mean free path. Since the thermal conductivity is directly proportional to the phonon mean free path, this results in reduced thermal conductivity for the thin film silicon in the fin (in a FinFET) or NW (in a GAAFET). In particular, since the dimensions of the fin/NW are of the same order of the mean free path, boundary scattering is exacerbated which has reduced the thermal conductivity of silicon by 10× in a fin and by 25× in a NW [7]. The NW has lower thermal conductivity since its dimensions are smaller than those of the fin. The larger surface to volume ratio of the NW leads to a larger material boundary thermal resistance, and greater phonon scattering at the perimeter of the NW.

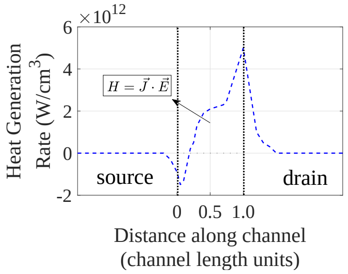


Figure 2: Profile of heat generation in the transistor channel [4].

We leverage the Joule heating equation from [4], where the rate of heat generation H (W/m^3) within the channel of a transistor is given by the dot product of the current density, $\vec{J}$, and the electric field, $\vec{E}$, within the transistor:

$$H = \vec{J}\cdot\vec{E} \qquad (3)$$

As current flows from drain to source in an NMOS device, the field is highest at the drain/drain extension [4], [8]. Therefore, spatially, the heat H, as given by (3), shows peak temperatures near the drain terminal of the device (Fig. 2). We use this to determine the spatial distribution of heat sources within the finite difference elements within the transistor channel.

C. Gate-level and Circuit-Level Thermal Analysis

In multigate FET technologies, gate layouts are highly regular, as illustrated for FinFETs by the array of horizontal fins crossing a set of vertical gate terminals in Fig. 1(a). For thermal analysis, we map these layouts into separate substructures depending on whether the transistors are in a series or parallel connection. We then apply a superposition-based methodology to compute the SH temperature in a circuit, consisting of a one-time *look-up table (LUT) precharacterization* of array substructures, followed by circuit-dependent analyses of *allocation of power dissipation to transistors* within the substructures and *LUT lookups* to obtain device temperatures.

Array precharacterization This is a one-time computation that can be considered to be a part of the library characterization. A set of fin/gate templates can be used to characterize all logic gates in a library. The template comprises of arrays with transistors either in a series connection or in a parallel connection with varying number of gate terminals (N), fins (F), and different options for Vdd/Gnd connections. The latter act as heat conduits through the power distribution network, especially in SOI/GAAFET circuits. The choice of connection points impacts the heat removal paths, and hence the temperature.

For each template, an FDM-based thermal simulation (Section II-A) characterizes the temperature for a unit power excitation, equally distributed in the channels of all fins/NWs controlled by the input, for each input of the logic gate. In general, we characterize two 3D LUTs (one for a series and the other for a parallel transistor configuration), each with a size of $N \times N \times F$. Each 2D $N \times N$ matrix in the LUT (Table I) provides values to estimate the temperature rise of a cell for a fixed number of fins. For example, Table I provides

Table I: LUT structure for a terminal array with 3 fins and up to 4 gate terminals in a parallel configuration for SOI FinFET technology.

Temperature rise (K) for a unit power excitation at each gate terminal for $F = 3$, $N = 4$, with contacts adjacent to gate terminals I_2 and I_3				
#	Current active input			
in	I_1	I_2	I_3	I_4
1	$T_{11}^{(1)}$	–	–	–
2	$T_{11}^{(2)}, T_{12}^{(2)}$	$T_{21}^{(2)}, T_{22}^{(2)}$	–	–
3	$T_{11}^{(3)}, T_{12}^{(3)}, T_{13}^{(3)}$	$T_{21}^{(3)}, T_{22}^{(3)}, T_{23}^{(3)}$	$T_{31}^{(3)}, T_{32}^{(3)}, T_{33}^{(3)}$	–
4	$T_{11}^{(4)}, T_{12}^{(4)}, T_{13}^{(4)}, T_{14}^{(4)}$	$T_{21}^{(4)}, T_{22}^{(4)}, T_{23}^{(4)}, T_{24}^{(4)}$	$T_{31}^{(4)}, T_{32}^{(4)}, T_{33}^{(4)}, T_{34}^{(4)}$	$T_{41}^{(4)}, T_{42}^{(4)}, T_{43}^{(4)}, T_{44}^{(4)}$

the temperature values of parallel configuration of up to 4 transistors for a fixed number of fins, $F = 3$. For a cell with a different number of fins, we index Table I in the third dimension (not shown here).

The values from these two 3D LUTs (one for the parallel and one for the series configuration of the transistors) can be used to estimate the temperature of structures with various combinations of serial and parallel sub-arrays.

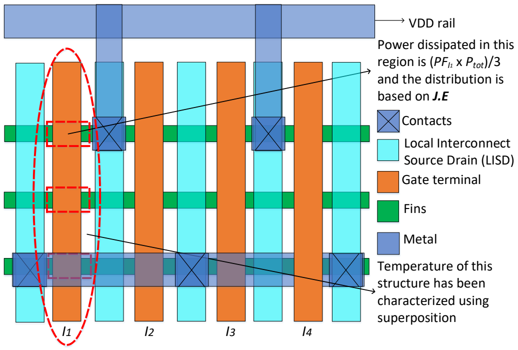

VDD rail
Power dissipated in this region is (PF_{I_1} x P_{tot})/3 and the distribution is based on ***J.E***
Contacts
Local Interconnect Source Drain (LISD)
Gate terminal
Fins
Metal
Temperature of this structure has been characterized using superposition
I1
I2
I3
I4

Figure 3: A sample 4x3 FinFET array substructure in a parallel configuration depicting the distribution of power within the array.

Example: Consider the template of parallelly-connected transistors in Fig. 3 with $F = 3$ fins and $N = 4$ gate terminals. Table I shows the 2D structure of the LUT for this figure, that characterizes this 4x3 array. A unit excitation in transistor i could cause an increase in temperature in a transistor j in the array: this is represented by $T_{ij}^{(k)}$, where $k \in 1, 2, ..N$, is the number of gates used in the current template.

Allocation of power to transistors in an array We now describe the distribution of the power dissipated in a gate, in its circuit context, to each input of the gate.

Standard power analysis methodologies can provide the switching and leakage power for a gate in its circuit context, based on the load capacitance and the activity factor. To determine how this gate-level power is allocated among the gate inputs, we consider an example of a parallel configuration of transistors with three inputs I_1, I_2, and I_3. The power dissipated while I_1 switches corresponds to four cases: I_2 and I_3 switch; one of I_2 and I_3 switches and the other is off; both are off. In these cases, I_1 carries a third, half, and all of the switching current, respectively. These lead to four terms used to compute the fraction of power for I_1, PF_{I_1}:

$$PF_{I_1} = \frac{\frac{\alpha_{I_1}\alpha_{I_2}\alpha_{I_3}}{3} + \frac{\alpha_{I_1}\alpha_{I_2}P_{I_3}}{2} + \frac{\alpha_{I_1}P_{I_2}\alpha_{I_3}}{2} + \alpha_{I_1}P_{I_2}P_{I_3}}{\alpha_Y} \tag{4}$$

Here α_{I_1}, α_{I_2}, α_{I_3}, and α_Y are the switching probabilities of inputs I_1, I_2, I_3 and output Y; P_{I_1}, P_{I_2}, and P_{I_3} are the probabilities that transistors I_1, I_2, and I_3 are off.

A similar relation can be derived for leakage power in a parallel stack, or for a series connection of devices. Since each device in the series connection have approximately similar switching resistances, and each of them needs to be ON for power dissipation, we distribute the net power equally among each of these devices. The net power is calculated based on the probability of all the devices in the series configuration being ON. This can be obtained in a similar manner as in (4).

LUT lookups to find device temperatures Given the allocation of power to each transistor in the array, we may now use the characterized LUTs to find the temperature in each transistor. Here, we can leverage the fact that thermal analysis involves solving a linear system and use the principle of superposition on (2).[1] Since the LUT entry $T_{ij}^{(k)}$ for the corresponding structure k provides the temperature rise in transistor j, with F fins, due to a unit power dissipation in transistor i, we compute the temperature in transistor j as

$$\text{Temperature}(j) = \textstyle\sum_i P_i T_{ij}^{(k)} \tag{5}$$

where $P_i = P_{tot} \times PF_i$ is the power dissipated in transistor i, computed as the product of the power dissipated in the gate and the power fraction for transistor i, and the summation is carried out over all transistors with N inputs in the array.

For example, if Table I is the LUT of the structure in Fig. 3, with $PF_{I_1} = 0.8, PF_{I_2} = 0.1, PF_{I_3} = 0.1$, and $P_{tot} = 0.1$, by using Eq. (5), we obtain a temperature rise in transistor I_1, with $k = 3$ gates used, as $(0.08T_{11}^{(3)} + 0.01T_{21}^{(3)} + 0.01T_{31}^{(3)})$.

III. Performance and Reliability Models

A. Effect of Temperature on Power and Delay

The temperature of the device affects both the delay and leakage power dissipation of the device. The subthreshold leakage, I_{sub}, current increases exponentially with temperature:

$$I_{sub} = I_0 \left(1 - e^{-\frac{V_{ds}}{V_T}}\right) e^{\frac{V_{gs} - V_{th} + \eta V_{ds} - k_\gamma V_{sb}}{nV_T}} \tag{6}$$

where all terms have their usual meanings, and the impact of temperature lies within the thermal voltage, $V_T = kT/q$, where k is Boltzmann's constant and q is the electron charge. Leakage-temperature feedback [6] is considered in an iterative manner during LUT characterization. However, since the threshold voltage degrades over time, the impact of leakage-temperature feedback is also reduced as the circuit ages, and the static power dissipation improves with time [9].

Increased temperature due to SH also impacts delays [6]. Due to increased scattering, the mobility of the carriers is degraded, increasing gate delays. However, the threshold voltage, V_{th}, also degrades, which has the opposite effect. This work uses SPICE-calibrated delay models that incorporate these factors.

B. Effect of Temperature on Reliability

We study the impact of SH on reliability; specifically, BTI [10] due to prolonged voltage stress on the gate, HCI [11], [12] due to the impact of carrier energy during transitions, and EM [13],which causes failures in wires due to metal migration. Degradations due to HCI and BTI are modeled as threshold voltage shifts over time, with different equations that reflect

[1] In practice, due to leakage-temperature feedback, the system is weakly nonlinear and may require a few iterations between leakage and temperature.

Table II: Reliability models used in this paper.

	BTI	HCI	EM
7nm Bulk FinFET	S. Mishra et al. [10]	Z. Yu et al. [11]	K-D. Lee [13]
7nm SOI FinFET		I. Messaris et al. [12]	
5nm GAAFET		Negligible [14], [15]	

the specific mechanisms, while the time-to-failure (TTF) for EM is commonly modeled by empirical equations. All three phenomena are accelerated at higher temperatures.

Table II summarizes the reliability models used in this work to analyze the impact of SH. In GAAFETs, the impact of HCI is minimal [14]: for 5nm GAAFETs, the time exponent n for HCI is 0.1 and is dominated by BTI [15]. Under BTI and HCI, the threshold voltage shift [16], due to the impact of temperature, $\Delta V_{th}(T)$, has the form:

$$\Delta V_{th}(T_{SH}) = \Delta V_{th}(T_0) e^{-\frac{E_a}{k}\left(\frac{1}{T_0} - \frac{1}{T_{SH}}\right)} \tag{7}$$

where T_0 is a baseline temperature for BTI/HCI modeling, and T_{SH} is the increased temperature due to self-heating.

For EM, the TTF shows a lognormal distribution, and the lifetime t_z for a target fail-fraction z is given by [13]:

$$t_z = t_{50} e^{z/\sigma} \tag{8}$$

where t_{50} is the mean time to failure (MTTF), and σ is the standard deviation of the lognormal failure distribution. The MTTF depends on current density J through the interconnect, and temperature T, and is modeled by Black's equation:

$$t_{50} = A J^{-n} e^{E_a/kT} \tag{9}$$

where $E_a = 0.45$eV is the activation energy for EM in Cu, n is the current exponent, and A is a fitting parameter.

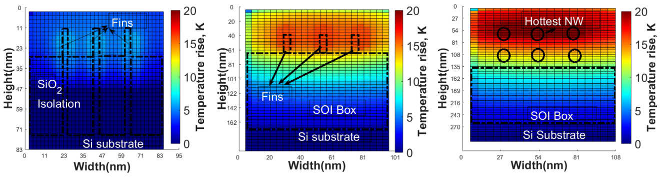
Due to the limited heat conduits to the ambient, the elevated device temperature is experienced in nearby wires. The device temperature thus directly influences EM in nearby wires. Very short wires may not be affected by EM due to the Blech criterion, which states that wires whose jL product (product of current density and length) is below a threshold. A Blech filteris typically applied before using Black's formula, and our EM results apply to wires that are Blech-mortal.

From (8) and (9), it is easily shown that the shift in t_z due to an SH-induced temperature shift from T_0 to T_{SH} has the same form as (7), but with $\Delta V_{th}(T_{SH})$ replaced by $\Delta t_z(T_{SH})$.

IV. Analysis of SH on FinFET Substructures

Table III lists the transistor dimensions, also illustrated in Fig. 1, for FinFETs [1] and GAAFETs [17]. Table IV lists the thermal conductivities of constituent materials [7], [8]: the values for fins/nanowires are lower than bulk due to phonon interactions. The paths to thermal ground from the FETs are (a) along the fin and then through the metal contacts and interconnects, and (b) downwards through the BOX (if any) and the substrate.

Comparing SH in various FET types: Fig. 4 shows the result of our thermal analysis approach on an array of 3 fins and 2 gate terminals, connected in series, of a cell from the ASAP7 library. It displays the cross-sectional temperature profile for bulk (Fig. 4 (a)) and SOI FinFETs (Fig. 4 (b)) and GAAFETs (Fig. 4 (c)). The contours show the qualitative difference between the three structures: the bulk FinFET, which has the easiest path to thermal ground has the lowest temperatures, followed by the SOI FinFET, where the bulk path is impeded by BOX, and then the GAAFET, where thermal paths must negotiate both BOX and the oxide surrounding the NWs. For the bulk FinFET, the substrate is the primary path to ground.



(a) (b) (c)

Figure 4: Temperature distribution for a power dissipation of 0.1μW for (a) a bulk FinFET with 3 fins/2 gates (b) an SOI FinFET with 3 fins/2 gates (c) a lateral GAAFET with 3 NW stacks/2 gates.

Table III: Physical dimensions of FinFETs and GAAFETs.

7nm Bulk/SOI FinFET		5nm lateral GAAFET	
Fin width	7nm	NW diameter	6nm
Fin height (Bulk/SOI)	32nm	NW horizontal spacing	14nm
Fin pitch	27nm	NW vertical spacing	10.6nm
Gate length	20nm	Gate length	12nm
Gate pitch	55nm	Gate pitch	33nm
Oxide thickness	1.8nm	Oxide thickness	1nm
SOI BOX thickness	140nm	SOI BOX thickness	140nm
Substrate thickness	1μm	Substrate thickness	1μm
Contact width/length	18mm	Contact width/length	10nm
Gate to contact	8nm	Gate to contact	4nm

Impact of contact locations: For the SOI FinFET, and much more so for the GAAFET, the path through the interconnects is a significant contributor to heat removal. We now examine how the locations of the supply contacts impact SH. As an example, we examine a NAND4X1 cell with 3 fins and 4 inputs for both SOI FinFET and GAAFET technologies. The inputs $I_1, \cdots, I_4$, have switching probabilities of 0.1, 0.1, 0.1, and 0.7, respectively. Fig. 3 shows the pull-up network (PUN) of this cell. The linear arrangement of PUN transistors (a parallel connection) can have two connections to V_{dd} between inputs: (i) I_1 and I_2, and (ii) I_3 and I_4 and three contacts to the other end of the parallel connection, as shown in Fig. 3. Another possible parallel configuration is having three connections to V_{dd} and the other two contacts shorted to complete the parallel connection. The pull-down network (PDN) is a series connection between the transistors and may be connected to ground at either input I_1 or I_4.

Table V shows the temperature rise for the PUN and PDN for each possible configuration. Since the contacts act as a heat conduction path, the PUN shows lower SH when three contacts to V_{dd} are used instead of two. For similar reasons, the temperature in the PDN is lower when the most active input, I_4, is closer to the ground contact. This leads to two guidelines: (i) standard cells should be designed to maximize the number of contacts to the supply network, and (ii) during circuit optimization, inputs that switch more often in a series chain should be placed closer to the supply. The latter recommendation, of course, must be balanced by the delay impact of pin selection (which may use arrival times to determine the proximity of inputs to the output or supply).

Impact of array size: When we increase the array size (number of fins or gate terminals) while keeping the power per input constant, the temperature in the array increases, as depicted in Fig. 5(a). Particularly, for SOI FinFETs and GAAFETs, where

Table IV: Thermal conductivites of the transistor materials [7], [8].

Material	Location	K (W/mK)	Material	Location	K (W/mK)
Si	Fins	13	SiO_2	SOI BOX	0.8
Si	Nanowires	5	HfO_2	Gate dielectric	0.27
Si	Substrate	148	Metal	Gate	48
Cu	Interconnect	42.4			

Table V: Temperature rise in a NAND4X1 cell for different contact location configurations of the PUN and PDN in both 5nm GAAFET and 7nm SOI FinFET technologies.

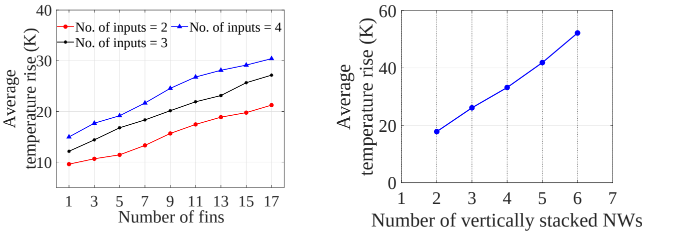
Temperature rise in PUN			Temperature rise in PDN		
Configuration	GAAFET	SOI	Configuration	GAAFET	SOI
2 contacts	18.1K	8.8K	Contact adjacent to I_1	31.4K	21.1K
3 contacts	13.8K	6.5K	Contact adjacent to I_4	23.3K	18.4K

the path through the BOX/substrate has high thermal resistance, the heat sinking paths through supply wires at the edge of the array are more distant in larger arrays. Thus, though we keep the total power unchanged, the effective resistance to thermal ground increases in the larger array. For a vertical GAAFET NW stack, similarly, an increase in peak temperature is seen as the number of vertically stacked NWs rises (Fig. 5(b)); here, upper NWs have a longer path to ground than lower NWs.

Impact of NW stack structure: The temperature distribution of a device with a single NW stack and multiple NW stacks is shown in Fig. 6. It is seen that the inner stacks heat up more than the outer stacks, owing to the fact that the inner stacks have higher surrounding temperatures. Similarly, for a FinFET, the inner fin is hotter than the outer fins or the single fin.

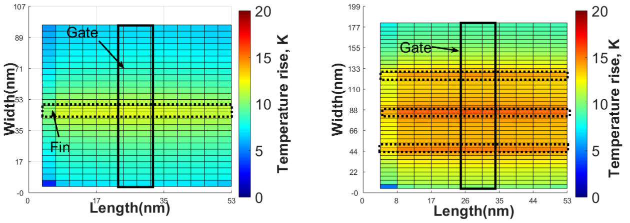
V. Circuit Level Self-heating Effects

We now consider the impact of SH on circuit delay and reliability, for various combinational ISCAS '85 and ITC '99 benchmarks. We synthesize the benchmarks using gates from the ASAP7 library [1], which includes NOT, 2–4 input NAND, AND, NOR, OR, AO, OA, AOI, and OAI gates. The circuit delay at different ages is calculated by performing static timing analysis (STA) with a frequency of 1GHz and a supply voltage of 0.7V, and for the analysis with SH, the LUT-based thermal model from Section II-C is integrated into the STA algorithm. The dynamic power dissipation for each gate is based on the load and its activity factor in the circuit, and the leakage power is taken from the characterized library [1]. Gate delay models are based on the standard non-linear delay model (NLDM) format, characterized at different temperatures and ages by running HSPICE simulations using the model files from the ASAP7 library [1]. The V_{th} increase due to HCI and BTI, and the EM lifetime, are modeled as in Section III-B.



(a) (b)

Figure 5: Temperature dependence on the (a) number of fins in a SOI FinFET and (b) number of vertically stacked NWs in a GAAFET.



(a) (b)

Figure 6: Temperature distribution for a GAAFET inverter with 1 gate terminal with (a) one and (b) three stacks of lateral NWs.

The average and peak temperature over all logic gates in each benchmark is shown in Fig. 7. Circuits that tend to show higher temperatures have a larger number of 4-input gates, with a higher resistance to thermal ground, and higher numbers of gates with high switching probabilities. These characteristics tend to be more visible in larger benchmarks. On an average, each gate in the benchmark heats up by 5K for bulk FinFET, 12K for SOI FinFET, and 18K for GAAFET technology.

These temperature distributions from the thermal analysis are used to estimate the impact on BTI, HCI, and EM. Fig. 8 shows the percentage degradation in delay shifts due to BTI and HCI after 10 years, while Fig. 9 shows the percentage lifetime degradation due to EM. As expected, the BTI and EM degradations correlate with the temperature rises, with the largest degradations for GAAFETs, followed by SOI FinFETs and then bulk FinFETs. For HCI on the other hand, despite the temperature of the bulk FinFET being lower, the degradation in bulk FinFETs is higher since the time exponent is larger than that of the SOI FinFETs. Degradations in SOI and GAAFET technologies are particularly large for EM, indicating that wider wires must be used for non-Blech interconnects in these technologies to ensure reliability. The magnitude of BTI and HCI shifts indicates that appropriate aging margins must be added in timing optimization to account for aging in emerging devices. For all the benchmarks shown, an average delay degradation of 10.3% for 7nm bulk FinFET technology, 12.4% for SOI FinFET technology, and 9.9% for 5nm GAAFET technology is observed due to SH. Despite the 5nm GAAFET having higher temperatures, the impact of HCI is dominated by BTI. The average delay degradation due to SH is thereby smaller in GAAFET compared to the other two technologies. On average, the percentage change EM-induced TTF is 14% for bulk FinFET, 38% for SOI FinFET and 45% for GAAFET.

The precise impact of SH on BTI and HCI, as a function of time, is shown in Fig. 10 for the b20 benchmark. Despite the impact of HCI being smaller in SOI FinFETs, the acceleration in aging in bulk FinFETs (28%) due to SH is lower than that for SOI FinFETs (37%). This is attributed to higher temperatures in that SOI devices, together with the impact of BTI. The value on each plot at time 0 is shown by dotted lines, and the difference between these dotted lines represents the delay degradation due to temperature alone, without aging (Section III-A), and is largely attributable to mobility degradation for this library. Overall, the plots show that SH in GAAFETs accelerate aging by an average of 62% over 10 years. It can be noted that for SOI FinFETs and GAAFETs, these SH-accelerated shifts can be quite significant even for low-lifetime parts (3–5 years).

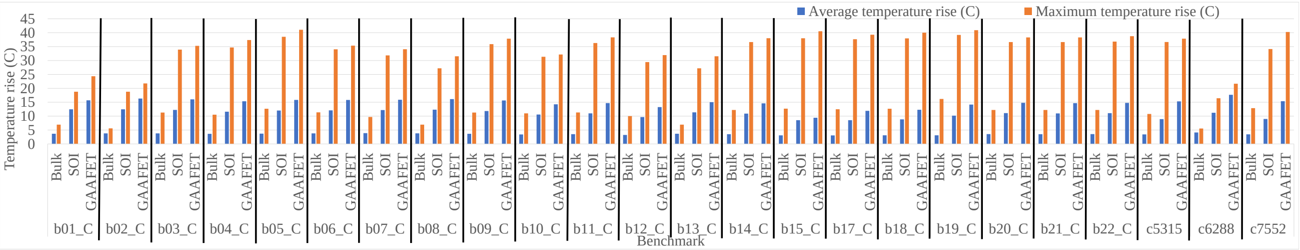


Figure 7: Average and peak temperature rises of different benchmarks for the different device architectures.

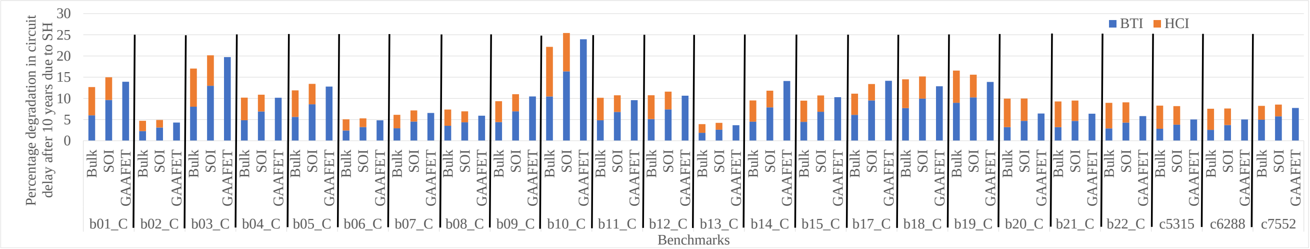


Figure 8: Impact of SH on BTI and HCI-induced circuit delay degradation after 10 years.

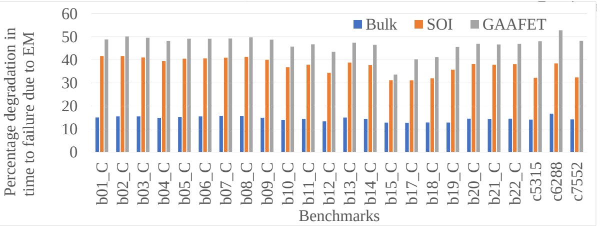
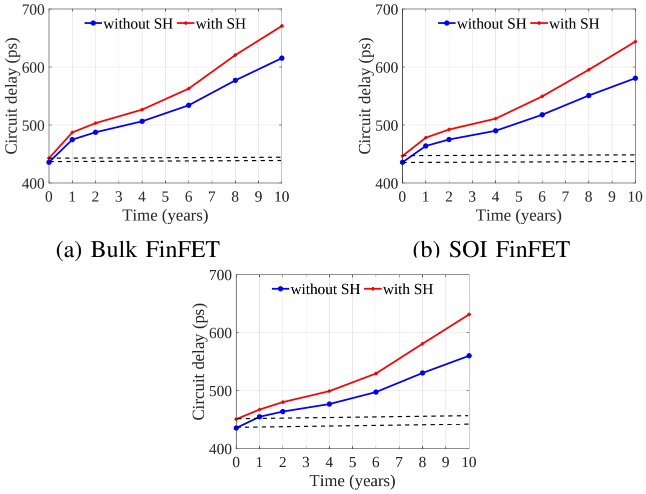


Figure 9: EM-induced time to failure, on benchmark circuits for bulk FinFETs, SOI FinFETs, and GAAFETs.



(a) Bulk FinFET

(b) SOI FinFET

(c) GAAFET

Figure 10: Circuit delay trends due to BTI and HCI aging over 10 years for the b20 benchmark for various technologies. At time 0, SH-induced aging changes mobilities and slows down the circuit. Over time, the rate of aging (the slope of red line vs. the blue line) increases due to SH.

VI. Conclusion

This paper has quantified the impact of modern multigate FET structures on SH, and its impact on performance and lifetime. In advanced structures, heat conduction paths face larger thermal resistances to the ambient and are prone to SH. This leads to the acceleration of aging mechanism, causing circuit behavior to degrade. Thermally-focused device optimization and circuit-level thermal management will be essential in the future to manage performance and reliability degradation.

References

[1] L. T. Clark, *et al.*, "ASAP7: A 7-nm FinFET predictive process design kit," *Microelectron. Reliab.*, vol. 53, pp. 105–115, 2016.

[2] Y. Huang, *et al.*, "GAAFET versus pragmatic FinFET at the 5nm Si-based CMOS technology node," *IEEE Journal of the Electron Devices Society*, vol. 5, no. 3, pp. 164–169, 2017.

[3] S. Ramey, *et al.*, "Aging model challenges in deeply scaled tri-gate technologies," *Proc. IIRW*, pp. 56–62, 2015.

[4] E. Pop, *et al.*, "Thermal analysis of ultra-thin body device scaling [SOI and FinFET devices]," in *Proc. IEDM*, pp. 36.6.1–36.6.4, 2003.

[5] S. E. Liu, *et al.*, "Self-heating effect in FinFETs and its impact on devices reliability characterization," in *Proc. IRPS*, pp. 4A.4.1–4A.4.4, 2014.

[6] Y. Zhan, *et al.*, "Thermally aware design," *Found. Trends Electron. Des. Autom.*, vol. 2, no. 3, pp. 255–370, 2008.

[7] E. Pop, *et al.*, "Heat generation and transport in nanometer-scale transistors," *Proc. of the IEEE*, vol. 94, no. 8, pp. 1587–1601, 2006.

[8] C. Xu, *et al.*, "Analytical thermal model for self-heating in advanced FinFET devices with implications for design and reliability," *IEEE T. Comput. Aid D.*, vol. 32, no. 7, pp. 1045–1058, 2013.

[9] D. Rossi, *et al.*, "Aging benefits in nanometer CMOS designs," *IEEE T. Circuits-II*, vol. 64, pp. 324–328, March 2017.

[10] S. Mishra, *et al.*, "TCAD-based predictive NBTI framework for sub-20-nm node device design considerations," *IEEE Trans. Electron Devices*, vol. 63, no. 12, pp. 4624–4631, 2016.

[11] Z. Yu, *et al.*, "New insights into the hot carrier degradation (HCD) in FinFET: New observations, unified compact model, and impacts on circuit reliability," in *Proc. IEDM*, pp. 7.2.1–7.2.4, Dec 2017.

[12] I. Messaris, *et al.*, "Hot carrier degradation modeling of short-channel n-FinFETs," in *Proc. DRC*, pp. 183–184, 2015.

[13] K.-D. Lee, *Electromigration critical length effect and early failures in Cu/oxide and Cu/low k interconnects*. PhD thesis, Univ. Texas Austin, Austin, TX, 2003.

[14] R. Wang, *et al.*, "New observations on the hot carrier and NBTI reliability of silicon nanowire transistors," in *Proc. IEDM*, pp. 821–824, 2007.

[15] M. Si, *et al.*, "Characterization and reliability of III-V gate-all-around MOSFETs," in *Proc. IRPS*, pp. 4A.1.1–4A.1.6, 2015.

[16] S. V. Kumar, *et al.*, "NBTI-aware synthesis of digital circuits," in *Proc. DAC*, pp. 370–375, 2007.

[17] H. Mertens, *et al.*, "Gate-all-around MOSFETs based on vertically stacked horizontal Si nanowires in a replacement metal gate process on bulk Si substrates," in *IEEE Symp. VLSI Technol.*, pp. 1–2, 2016.